BIOINFORMATIC PLATFORMS AND METHODS FOR WORLDWIDE POLYGENIC RISK SCORES

A Dissertation
Presented to
The Academic Faculty

by

Aroon T. Chande

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
December 2020

BIOINFORMATIC PLATFORMS AND METHODS FOR WORLDWIDE POLYGENIC RISK SCORES

Approved by:

Dr. I. King Jordan, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Joseph Lachance
School of Biological Sciences
Georgia Institute of Technology

Dr. Gregory Gibson
School of Biological Sciences
Georgia Institute of Technology

Dr. Augusto Valderramma-Aquirre
Faculty of Health
Universidad Santiago de Cali

Dr. Soojin Yi
School of Biological Sciences
Georgia Institute of Technology

Date Approved: July 30, 2020

In dedication to Dr. Shashikala Sukhatme – a brilliant statistician, loving grandmother, and my most ardent supporter.

ACKNOWLEDGEMENTS

My growth and development as a researcher and bioinformatics scientist would not have been possible without the guidance of my advisor, Dr. I. King Jordan, and his fanatical pursuit of new scientific frontiers. King's heartfelt desire to inspire and encourage his students to find the research they are passionate about was a great motivation throughout my graduate career. I hope I can keep a similar passion for science throughout my own career. I am also thankful for his patience, understanding, and generosity of praise; little things make a big difference.

I would also like to thank the members of my committee – Dr. Soojin Yi, Dr. Greg Gibson, Dr. Joe Lachance, and Dr. Augusto Valderrama-Aguirre – for their support and guidance. The expert eyes and suggestions each of you brought to my research was invaluable. I truly appreciate the broader perspective offered by all of you; it is all too easy to miss the forest for the trees.

I must also thank all my friends and colleagues at the Applied Bioinformatics Laboratory and in the Jordan lab and School of Biological Sciences. I cannot overstate the impact you have all had on my life over the last five years. Lavanya Rishishwar and Michael Astwood have my deepest gratitude for their belief in my abilities and their guidance throughout the years. ABiL gave me great freedom to explore and grow as a scientist while working across a wide range of disciplines. Much of my work would not have been possible without support from Troy Hilley and his IT wizardry.

Finally, this entire endeavor would not have been possible without the love and support of my family and friends. To my parents, Drs. Vidya and Tushar Chande, and brother Ravi, I owe a debt of gratitude for their unwavering love and support. Their encouragement in hard times and pride in my achievements kept me going. There is no better pick me up for a Ph.D. student than a box full of food, lovingly made by their mom, dad, and brother from across the country. And to my Atlanta family – Anna Gaines, Jose Jaimes, Barry Miller, Shashwat Deepali Nagar, Sharllyn Pimental, and Devika Singh – I could ask for no better friends, I love you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1KGP	1000 Genomes Project
AF	Allele frequency
AFR	African
CAD	Coronary artery disease
CCD	Common complex disease
CGPB	Colombian Genotype-Phenotype Browser
CLM	Colombian in Medellin, Colombia
DAF	Derived allele frequency
EA	European-American
EAS	East Asian
EUR	European
F_{ST}	Fixation index
GADGET	Global Distribution of Genetic Traits webserver
GWAS	Genome-wide association study
HDI	Human development index
HL	Hispanic/Latino
LA	Latin America
LD	Linkage Disequilibrium
MA	Mexican-American
MAF	Minor allele frequency
MsigDB	Molecular Signatures Database
MXL	Mexican Ancestry in Los Angeles, California
OR	Odds ratio
PCA	Principal component analysis
PGS	Polygenic score
PRS	polygenic risk score
PTS	Polygenic trait score
SAS	South Asian
SES	Socioeconomic status
SNP	Single Nucleotide Polymorphism
T2D	Type 2 diabetes
uPTS	Unweighted PTS
wPTS	Weighted PTS

SUMMARY

This thesis evaluates the effects of ancestry and admixture on the utility and generalizability of polygenic scores for risk prediction and stratification in diverse populations in the US and Latin America. Historically, biomedical research has focused predominantly on European-origin individuals to the neglect of ancestrally and ethnically diverse populations. My work is motivated by health disparities that disproportionately affect disadvantaged populations across the world, and in particular, those in the Americas. New World populations from the Americas are characterized by varying degrees of admixture among ancestral population groups from Africa, the Americas, and Europe. I utilize thousands of genomes from diverse populations worldwide, along with hundreds of genome-wide association studies (GWAS) on thousands of human traits, to address three overarching questions: (1) which phenotypes vary among populations, and what explains that variance?; (2) is it possible to predict and stratify risk for common complex diseases across diverse populations?; and (3) can we apply already discovered genetic associations to risk prediction in new and ancestrally distinct populations?.

In precision medicine, a patient's genetics are used to inform treatment decisions. This approach relies on knowledge about the genetic architecture of disease, along with environmental and lifestyle effects, to accurately predict outcomes. The treatments recommended by precision medicine can range from non-invasive preventative measures to highly invasive interventional practices – e.g., more frequent mammographic screening versus mastectomy in patients with predicted breast cancer risk. Therefore, it is vitally important that medical practitioners have the best and most complete knowledge about a

disease as possible. Large-scale population genetics studies, GWAS, are what enable this knowledge gathering. Moreover, precision public health seeks to apply precision medicine at the local and regional population level, which supports the implementation of public health initiatives that are most appropriate and most impactful for a given region.

The populations of many countries, with notable exceptions, are composed of admixed individuals with ancestry derived from two or more ancestry groups – African, Southeast and East Asian, European, and Native American. Two notable examples to the contrary, Iceland and Finland, are markedly homogenous – show the lowest between and within-population diversity of any populations studied in this dissertation. This homogeneity has made these two populations popular for genomics research; their shared life histories and societal influences greatly simplify genomics research – freeing researchers of confounding socioeconomic, historical, and genetic factors that would otherwise complicate their work. This theme of using simpler to study, ancestrally homogenous Eurocentric populations holds in the US and Latin America and has resulted in a large gap in genomics knowledge that only compounds existing health disparities.

Health disparities are "preventable differences in the burden of disease, injury, violence, or opportunities to achieve optimal health that are experienced by socially disadvantaged populations" [1]. In terms of genetics research, these disparities are exacerbated by a lack of inclusive research that studies individuals and genomes of diverse ancestry. This genomics research gap grows larger when it comes to translating research findings into the clinic as part of precision medicine initiatives. Therefore, we must find a way to integrate and apply our existing knowledge to the treatment of disadvantaged and currently underrepresented populations. My thesis research is focused on developing

platforms and methods that democratize access to population genetics research and support the application and analysis of existing information in new and diverse populations. This was accomplished through the development of online platforms for population genetics research and the assessment and development of risk prediction methods in diverse populations.

GWAS have helped elucidate the population-specific genetic architecture of diseases by associating particular single nucleotide polymorphisms (SNPs) with changes in measurable phenotypes. Polygenic scores (PGS) are an emerging translational genetics tool for predicting the magnitude of an individual's phenotype, for both anthropometric (such as height or body mass index) and disease-related traits. The standard approach to PGS construction is to use the results of a GWAS and related individual-level phenotypes data to iteratively arrive at the most informative set of SNP associations. However, this approach is limited to the population from which the GWAS sampled from and frequently produces population-specific PGS. My work attempts to generalize PGS into diverse populations and assess their effectiveness from a precision public health perspective.

My dissertation leverages GWAS, PGS, and epidemiology data to assess the utility of PGS for risk stratification of diseases in admixed Hispanic/Latino and Afro-descendant populations in the United States and Colombia. I focus on common complex diseases, which are polygenic disorders with high prevalence and multifactorial etiology, such as type 2 diabetes (T2D) and coronary artery disease (CAD), which, together, affect upwards of 80% of some populations. Relative differences in population PGS distributions are used to assess the effectiveness of PGS and controls developed herein. Finally, I explore the

relationship between ancestry and predicted outcomes across thousands of traits and evaluate the extent to which ancestry biases PGS.

Research advance 1: My first study explores the distribution of predicted phenotypic differences in 27 global populations, and I developed the first of its kind web-based population genetics platform for computing and visualizing PGS. The computational requirements for population genetics are frequently a substantial barrier to entry, particularly for researchers without bioinformatics expertise and from lower-income areas. Many of these lower-income areas are home to traditionally underserved populations, which are also uniquely at risk for particular complex common diseases [2, 3]. Web-based platforms have low barriers to entry, requiring no computational resources beyond an internet-connected device. I developed two platforms for exploring the relationship between global genetic diversity and the genomic architecture underlying human phenotypic diversity.

Research advance 2: My second study builds on the population precision health paradigm by integrating population genotype data with fine-scale epidemiology data for two adjacent and ancestrally diverse departments in Colombia. This study assessed the utility of PGS, without individual-level phenotype data, for stratifying population risk for three categories of disease with high socioeconomic impact: (1) common complex: T2D, hypertension, chronic kidney disease, and ischemic stroke; (2) cancers: cancers of the central nervous system, leukemia, Hodgkin and non-Hodgkin's lymphomas, ovarian, and pulmonary cancers; and (3) malaria caused by *Plasmodium vivax* and *P. falciparum*. Using PGS derived from multi-ethnic associations, I show that predictions tend to be highly correlated with outcomes at the population level. This study also highlights the impact of

putative environmental effects on confounding PGS interpretation. Genetic risk does not exist in a vacuum, and Genetic × Environmental effects must be accounted for in complex traits.

Research advance 3: My third and fourth studies focus on risk prediction for a single disease in two regional populations: type 2 diabetes in Hispanic/Latino populations in the US and Afro-descendant populations in Colombia. T2D has high prevalence across the globe, and disproportionately affects non-European ancestry populations throughout the western world. Diabetes is also one of the most and best-studied diseases in genetics, with highly diverse study populations compared to all other traits studied by GWAS. This diversity in existing research provided a unique opportunity to develop ancestry-specific and ancestry-agnostic PGS for T2D in four populations; Afro-Colombians in Chocó, Colombia, Mestizos in Medellin, Colombia, Mexican-Americans in LA County, California and European-Americans in Utah. The predicted risk for T2D was uniformly high for Afro-Colombians and Mexican-Americans, regardless of the GWAS and PGS ancestry composition. Strikingly, while low socioeconomic status is strongly correlated with higher T2D risk and prevalence in the US, in Colombia, it has a protective effect. Finally, a derived allele frequency-based control was proposed that enables the normalization of PGS across populations and ancestries. Preliminary results from this study suggest this control is effective.

CHAPTER 1. INTRODUCTION

Genetic diversity underpins much of the observed human phenotypic diversity and plays an important role in human health and disease [4]. As ancient human populations migrated and became reproductively isolated, they accumulated population-specific variations, some of which were advantageous in their respective environments. Eventually, these adaptive alleles increased in frequency and helped ancient communities survive by bolstering resistance to infectious disease [5, 6], aiding in the metabolism of complex biomolecules [7], and overcoming other environmental obstacles [8, 9]. In contemporary populations, these ancient adaptations may no longer be beneficial in the current environment and, through antagonistic pleiotropic effects, increase the risk of common complex diseases [10-12]. Common complex diseases are non-infectious polygenic disorders caused by the complex interaction between many genes and environmental factors, such as type 2 diabetes (T2D) and coronary artery disease (CAD) [13-15].

Genome-wide associations studies (GWAS) have helped elucidate the population-specific genetic architecture of diseases by associating particular single nucleotide polymorphisms (SNPs) with changes in measurable phenotypes. Thousands of GWAS, studying millions of people, have been performed over the last 15 years, but the majority – nearly 80% – of study participants have been of European ancestry [16]. As a result, there is a strong foundation of knowledge on the genetic architecture of disease in Europeans but not for ancestrally diverse populations in the Western world where most biomedical research takes place. While global populations share significant portions of disease architecture [17-20], the actual magnitude and direction of effects are often different [18, 21-23]. The reliability of applying GWAS knowledge ascertained in Europeans to populations of diverse ancestry remains an open question [16, 18, 22, 23].

One proposed, and likely successful, method to close this gap is more GWAS in diverse populations. Over the last ~10 years, researchers have increasingly pointed to the lack of genomic diversity and lobbied for more representation in studies. Indeed, the raw number of Afro-descendent, Hispanic/Latino, Native American, and Southeast Asian individuals represented in GWAS has increased dramatically; however, they still only represent a small portion of the total number of individuals studied (<20%) [24] despite making up the majority of peoples around the world. This is further compounded by re-analysis efforts on large Eurocentric biobank cohorts, which lead to an inflation of low-diversity GWAS. This effectively results in <10% of GWAS participants being from traditionally underrepresented populations.

However, performing high power and ethnically diverse GWAS is prohibitively expensive, with costs ranging into the several millions of dollars per study. More and larger GWAS in every population for every disease is not a sustainable or scalable approach. Therefore, we must also find a way to apply existing knowledge to new populations. The first steps in utilizing this existing knowledge are (1) estimating the transferability of European-derived SNP-trait associations into diverse populations, and (2) assessing and controlling for biases introduced while transferring this knowledge. Throughout this dissertation, these two themes will be discussed at length, with applications in Hispanic/Latino and Afro-descendant populations.

1.1 Polygenic traits and genome-wide association studies

Early genetics research, in both humans and other model organisms, focused on readily observable and reliability inherited phenotypes. These phenotypes are caused by genetic variants, or alterations in the genome, with the changes ranging from a single nucleotide change (SNPs) to the deletion of entire chromosomes. These variants, or alleles, tended to be both very rare ($AF \leq$

0.001) and highly penetrant. Penetrance is the probability of phenotypic expression given the presence of a phenotype-conferring allele. Highly penetrant alleles almost always confer phenotypic expression. As an aside, it is essential not to conflate penetrance, the mere expression of a phenotype, with expressivity, the strength or intensity of phenotypic expression. In section 1.2 and 1.3, I will discuss in more detail the relationship between penetrance, expressivity, and what PGS attempt predict. Diseases caused by such highly penetrant, rare alleles are often referred to as Mendelian diseases (Figure 1) since their mode of inheritance closely matches those noted by Gregor Mendel in his peas. Because these disorders are caused by a single mutation impacting a single gene, they are sometimes also called monogenic disorders.

On the other hand, common alleles ($AF \geq 0.1$) associated with disease tend to be low penetrance (Figure 1). A single low penetrance allele is unlikely to cause an observable phenotype, however in combination with few to many related low penetrance alleles, it can elicit a phenotype. This combination of alleles generally involves alleles that change the expression or function of multiple genes. Such phenotypes are polygenic and, in some cases, such as height, are considered omnigenic – they involve the action of nearly all genes [25].

Discovering and characterizing the alleles that influence polygenic phenotypes is the purview of the case-control association study, also known as a genome-wide association study (GWAS). In GWAS, many individuals who do (case) and do not (control) display a phenotype (e.g., diabetes) first have their genomic variation quantified. The most popular and cost-effective method of quantification is microarray genotyping, which accurately determines sequence variants at many predetermined sites along the genome. Investigators next scan along the genome and ask, "is the allele frequency of this variant significantly different between cases and controls?". In actual practice, performing a GWAS involves several controls, which are discussed below, and are

modeled in linear or logistic regression as covariates. GWAS studies may uncover protective alleles, alleles that are in significantly higher frequency in controls, and risk alleles, alleles that are in significantly higher frequency in cases. GWAS are not limited only to binary traits like disease status; they can also be applied to quantitative traits such as height or blood pressure. In such studies, the genotype frequency of each allele is linearly regressed with quantitative measures and other covariates, such as ancestry and lifestyle factors.

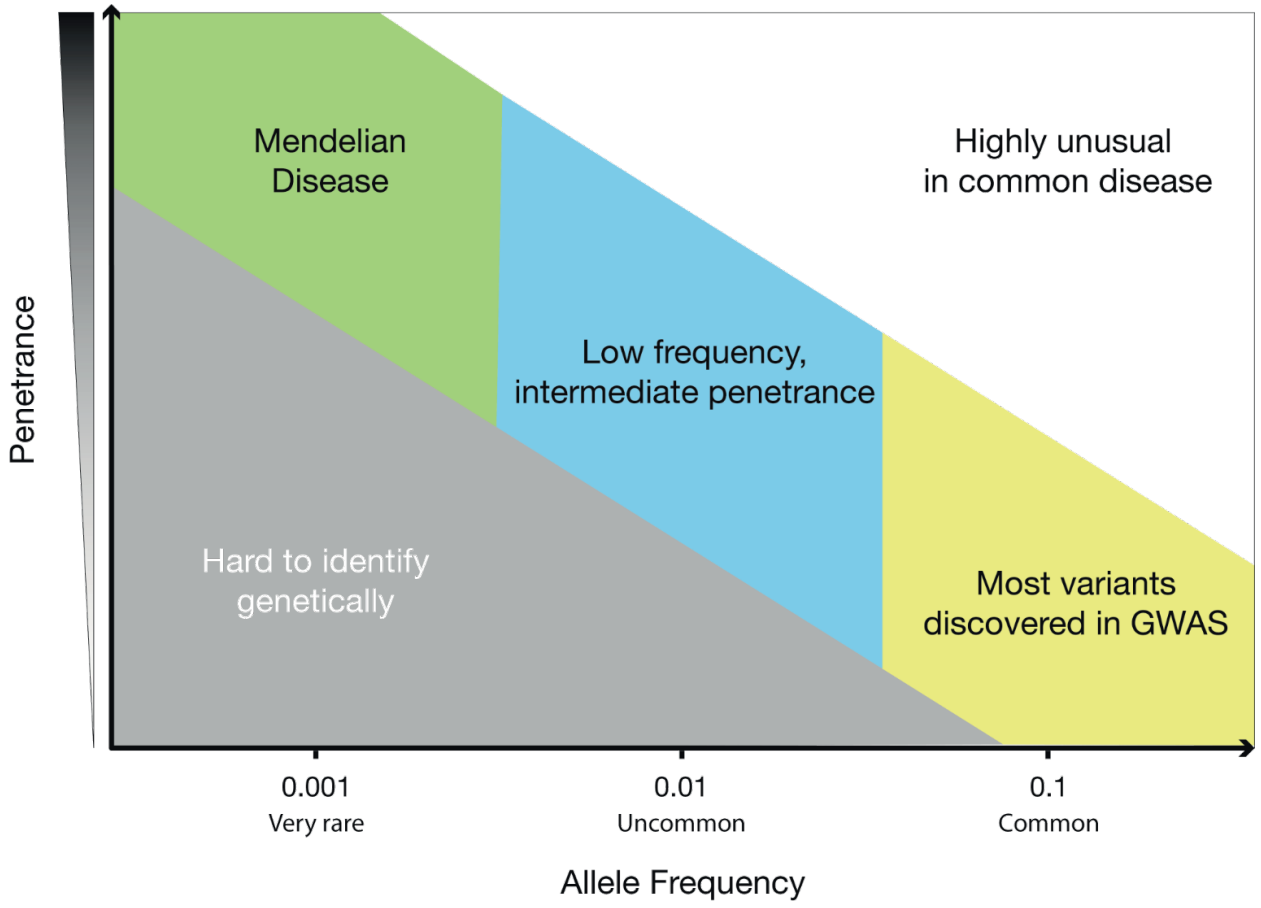

Figure 1. Allele frequency, penetrance, and disease

The relationship between allele frequency and penetrance, as it relates to genetics research and our ability to detect such associations. Adapted from [26].

Crucially, GWAS must correct for both genetic and non-genetic effects, which may introduce error in the results. Genetic effects include sample relatedness and population structure. Controlling for sample relatedness is often as simple as removing individuals who appear, based on their combination of genotypes, to be too closely related – in practice, this typically means eliminating individuals who appear to be second-degree relatives or closer. It is important to remove related individuals as they may have shared family-specific deviations in allele frequency that are unrelated to the phenotype of interest but skew the allele frequency differences tested. Correcting for population structure involves computing the relatedness (or distance) between all study samples, by performing principal components analysis or multi-dimensional scaling, and using the resulting component vectors as covariates in the association test. Of note, as shown in Figure 21, these components correspond to allele frequency differences between and within populations of different ancestry. Non-genetic effects include environmental exposures, such as smoking tobacco and drinking alcohol, and quantitative measures such as blood iron levels or waist-to-hip ratio. That smoking tobacco has a direct causal link to head and neck cancer is unsurprising in light of what we know about the carcinogenic compounds in cigarette smoke. However, we want to ascertain the genetic risk in the absence of the effects of tobacco smoke and include smoking status as a covariant in our association test model.

Regardless of the type of trait studied, the outputs of a GWAS are the per-variant minor allele frequency (MAF), odds ratio (OR, the ratio of non-reference allele frequency between conditions) or beta (β), standard error, and p-value. The widely-used value for genome-wide significance after multiple testing correction is $p < 5 \times 10^{-8}$, which was determined by simulation studies in European populations [27]. The odds ratio/beta values are used to determine

the allele effect, with OR < 1 indicating the allele is found at higher frequency in cases and increases (risk of) the phenotype.

1.2 Common complex diseases and public health

A common complex disease (CCD) is, as the name would suggest, commonly prevalent in a population and has a complex etiology. Estimates of commonness for CCD range from several percentage points, as is the case for many cancers, to upwards of 80% of the population in the case of diabetes and pre-diabetes metabolic disorder. CCD are polygenic disorders caused by the combination of many, low-penetrance variants across multiple genes. The hallmarks of CCD are the complex interaction between genetics and environment, so-called $G \times E$ interactions, and non-Mendelian inheritance. That risk for CCD is inheritable may seem obvious since the diseases arise from genetic variants but consider the case of cancers, which are driven by an accumulation of somatic mutations that cannot be passed to the next generation. Therefore there might exist a combination of heritable alleles that contribute to the accumulation of somatic mutations that have direct and indirect functional consequences on oncogenesis. GWAS enables researchers to discover both, directly and indirectly, risk modifying alleles and explore the environmental exposures that influence expressivity.

Precision medicine practitioners and public health agencies are particularly interested in CCDs since their expressivity is modifiable by environmental and lifestyle factors. Take, for example, the leading causes of morbidity and mortality in the US: heart disease and diabetes. The genetics and environmental mediators of both diseases are well studied across the US, in multi-ethnic regional populations. At the individual level, doctors can order panel-based genotyping tests to measure the genetic component of risk and recommend lifestyle changes to lower the

severity and risk of developing heart disease or diabetes. At a population level, using population genetics data, public health agencies can recommend lifestyle changes for large groups of people to help lower their risk without incurring the (massive) expense of genotyping every individual. A significant roadblock to this population precision medicine approach is the availability of appropriate ancestry-matched population genetics data. My dissertation attempts to expand the reach of these population precision medicine techniques by assessing the generalizability of existing population genetics knowledge across ancestries and assessing the utility of these predictions on populations in the US and Colombia.

Precision public health is "providing the right intervention to the right population at the right time" [28]. The goals of precision public health are to decrease population-specific health disparities. This is done through the targeted application of resources and technology, such as the recruitment and study of representative cohorts from communities, states, and across the country for genetic studies. The end goal is to enable genomics-based treatment guidelines without requiring all individuals to undergo expensive sequencing-based tests. In order to affect precision public health, we must understand population-specific risk profiles and be able to predict genetic risk.

1.3 Genetic risk prediction and stratification

An emerging tool in the precision medicine toolkit is the prediction of genetic predisposition and risk for disease using polygenic scores (PGS). PGS leverage the genetic association and effect size estimates from GWAS to predict the heritable phenotypic expression. The computation and correction of PGS will be covered in detail throughout my dissertation. Population PGS distributions are essential in understanding population-specific risks and

disparities. Population PGS distributions are useful in determining if any populations are uniquely at risk for, or protected from, a particular disease. This stratification of population risk into low, moderate, and high make PGS well suited for preventative efforts. Precision public health initiatives can focus on funding and awareness campaigns for high cost and high impact diseases that are predicted to be at high risk.

Despite their apparent potential for discovering genetic changes that underlie phenotypic divergence among populations, recent studies have underscored several challenges entailed by cross-population comparisons of PGS [29]. Systematic differences in allele frequencies, proportions of ancestral versus derived alleles, and patterns of linkage disequilibrium can yield significant shifts in PGS distributions that do not necessarily reflect observed phenotypic differences among populations [23, 30, 31]. Furthermore, the fact that the vast majority of GWAS have been conducted on cohorts of European ancestry [16, 32, 33] results in PGS that are far more accurate for European populations compared to other, less-studied global population groups [24]. In light of these challenges, my dissertation (1) characterize the genetic ancestry patterns for diverse populations from within two countries, (2) evaluates the impact of ancestry differences between these populations on the genetic variants associated with phenotype and function, and (3) consider observed differences in the frequencies of trait-associated variants in light of observed differences between the populations.

1.4 Genetic admixture and ancestry

Genetic admixture (herein "admixture") is the combination of alleles from two or more genetically differentiated populations. In its simplest form, admixture occurs when two individuals from previously geographically or reproductively isolated populations mate and have

offspring. These offspring have combinations of alleles, found in contiguous chromosomal blocks called haplotypes, that come from each parent. In the modern era, admixture can be much more complicated, especially when two admixed individuals mate. The new combination of alleles produced by admixture contains alleles adaptive for multiple environments, and whether or not they confer a selective advantage in the current environment is unknown.

In the above example, the allele frequencies within each haplotype are population-specific since they, by definition, must come from one parent or the other and are passed down through generations. As a result, we can computationally reconstruct ancestral haplotypes in a population with a high degree of certainty using allele frequency alone. Haplotype reconstruction becomes more difficult the older the admixture event(s) is because the haplotype structure breaks down as a result of genetic drift, adaption, and additional mating. We define genetic ancestry as the proportions of the genome that are attributable to one or more putative ancestral populations. Put simply, for each haplotype in an individual's genome, we find the ancestral population that has the best matching combination of alleles at the same genomic location, creating a block of ancestry. After this procedure, we sum up the length of each ancestry block to find the percent ancestry for each putative ancestral component in an individual's genome. The exact method of assigning ancestry varies and can produce either global or local ancestry estimates. Global estimates, produced by the program ADMIXTURE, do not reconstruct an individual's haplotypes, and thus does not directly attribute ancestry to given genomic loci but instead provided a single estimate across the entire genome for a given ancestry. In contrast, local ancestry, assigned by RFMix, involves the reconstruction and attribution of ancestry to specific genomic loci and can, therefore, be used to analyze patterns in admixture across the genome.

Estimates of global and local ancestry from ADMIXTURE and RFMix are highly concordant, and for most of the analysis in my dissertation, I will use global estimates. ADMIXTURE provides a K-way vector of K putative ancestral fraction that sum to ≈1, with K being a user-defined parameter defining the number of a priori expected ancestral populations. ADMIXTURE infers the ancestral groups based on allele frequencies within the input individuals by partitioning haplotypes into K groups. ADMIXTURE was run with K=3 ancestral populations for all the populations studied here, with the three populations corresponding to putative ancestral populations composed of African, European, and Asian/Native American individuals.

CHAPTER 2. EXPLORING THE WORLDWIDE DIVERSITY IN POLYGENIC TRAITS

2.1 Abstract

Human populations from around the world show striking phenotypic variation across a wide variety of traits. GWAS are used to uncover genetic variants that influence the expression of heritable human traits; accordingly, population-specific distributions of GWAS-implicated variants may shed light on the genetic basis of human phenotypic diversity. With this in mind, I developed the GlobAl Distribution of GEnetic Traits web server (GADGET http://gadget.biosci.gatech.edu). The GADGET web server provides users with a dynamic visual platform for exploring the relationship between worldwide genetic diversity and the genetic architecture underlying numerous human phenotypes. GADGET integrates trait-implicated single nucleotide polymorphisms (SNPs) from GWAS, with population genetic data from the 1000 Genomes Project, to calculate genome-wide polygenic trait scores (PTS) for 818 phenotypes in 2,504 individual genomes. Population-specific distributions of PTS are shown for 26 human populations across 5 continental population groups, with traits ordered based on the extent of variation observed among populations. Users of GADGET can also upload custom trait SNP sets to visualize global PTS distributions for their traits of interest.

2.2 Background

All human traits that have been measured thus far show evidence for some amount of heritability [4]. The expression of heritable traits is influenced, to varying degrees, by the presence of specific genetic variants. Since the frequencies of most genetic variants are known to vary

among human populations, heritable traits may be expected to differ across populations as well. Indeed, human populations around the world show tremendous variation for a wide variety of heritable traits.

GWAS can shed light on the genetic architecture underlying heritable human traits. Over the last ten or so years, numerous GWAS studies have been used to discover thousands of genetic variants that influence the expression of hundreds of human traits, including anthropomorphic, behavioral, and health-related phenotypes [34]. Exploration of the distribution of GWAS implicated variants across global populations has the potential to yield insight into the genetic basis of human phenotypic variation.

Heritable human traits are complex and polygenic; they are influenced by the action of genetic variants at multiple loci throughout the genome, along with environmental effects. Recently, genome-wide PGS have emerged as a powerful tool for predicting individuals' phenotypes based on the numbers of effect (risk) alleles encoded in their genomes [35-37]. PTS can be computed by summing the numbers of effect alleles encoded in an individual genome, and scores can be weighted by considering allele effect sizes. In the case of health-related phenotypes, PTS are often referred to as genetic risk scores, reflecting the predicted health risk to individuals entailed by the presence of disease-implicated variants in their genomes. We reasoned that the calculation of PTS for different human population groups could be used to shed light on population-specific variation for heritable human traits. To this end, we developed the GlobAl Distribution of GEnetic Traits (GADGET) web server, providing users with an intuitive tool for exploring the relationship between worldwide genetic diversity and the genomic architecture underlying a wide variety of human phenotypes (Figure 2). The GADGET web server allows users to explore the population-specific distributions of pre-computed PTS for >800 human traits

across 26 global populations. Users also have the option to upload custom SNP sets in order to assess the global distribution of PTS for their traits of interest.

It should be noted that GADGET is intended as a tool for researchers to explore population-specific distributions of genetic variants that have been associated with a wide variety of human traits. Users of the site should treat the results with caution, as the interpretation of PTS across populations can be complicated by many factors [31]. In this sense, the PTS distributions returned by GADGET can perhaps best be considered as working hypotheses, rather than definitive assertions of population-specific differences in genetic traits.

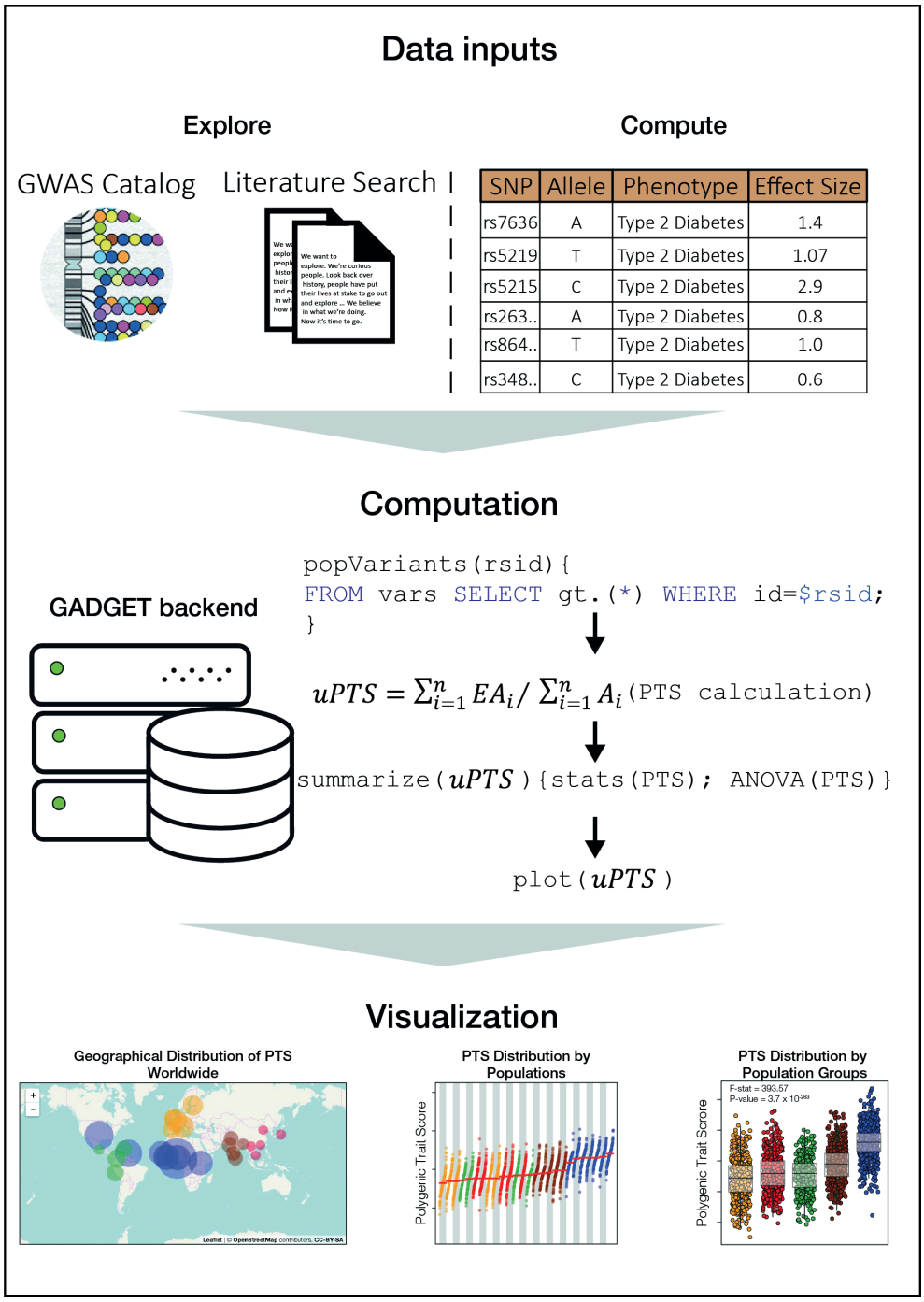

Figure 2. Schematic overview of the GADGET web server workflow.

Curated (Explore mode) or user-provided (Compute mode) trait SNP sets are used to calculate genome-wide PTS. Population-specific PTS distributions are shown for 26 global populations organized into 5 continental (super) population groups. Among population PTS variation is quantified with ANOVA (F-stats and P-values provided).

2.3 Material and Methods

2.3.1 GADGET Platform

The GlobAl Distribution of GEnetic Traits (GADGET) webserver frontend and visualizations are built using the R programming language (https://www.r-project.org) with the Shiny web development package (http://shiny.rstudio.com). The user-editable spreadsheet is provided by rhandsontable (https://github.com/jrowen/rhandsontable), and shinyjs (https://github.com/daattali/shinyjs) provides the R-to-javascript interface for custom tab interactions and input validation. GADGET results are visualized with ggplot2 (http://ggplot2.org) (box plots and strip charts) and leaflet.js (http://leafletjs.com) (maps). The computation backend consists of an SQL database and companion Perl and Python scripts for data validation and PTS score calculation. In an effort to simplify the deployment of the Explore module, pre-computed results are exported to flat-files. GADGET supports the sharing of results via the "Share Table" function.

2.3.2 Polygenic trait SNP data

The GADGET web server allows users to (1) explore the global distributions of pre-computed genome-wide polygenic trait scores (PTS) for >800 phenotypes, or (2) to compute worldwide PTS for user-defined traits of interest. The pre-computed PTS are based on trait-specific sets of SNPs taken from the NHGRI-EBI GWAS Catalog. In contrast, the user-defined PTS are derived using user-supplied sets of SNPs corresponding to traits of interest. Both pre-computed and user-defined PTS are calculated using genome sequence variant data for 2,504 individual genomes from 26 global populations characterized as part of the 1000 Genomes Project (1KGP).

Polygenic trait and SNP (effect allele) information used to compute PTS were taken from the NHGRI-EBI GWAS Catalog version 1.0.1 [34], and trait descriptors were organized into functionally coherent trait categories using the EBI Experimental Factor Ontology [38] November 2017 data release. We rely on the NHGRI-EBI GWAS Catalog study eligibility criteria and SNP reporting methods (https://www.ebi.ac.uk/gwas/docs/methods) to curate SNPs associated with specific traits; accordingly, SNPs are incorporated into our trait-specific SNP sets if they show a genome-wide association *P*-value < $1.0e^{-5}$. The trait-associated SNPs curated from the NHGRI-EBI GWAS Catalog are all based on dbSNP build 150 and the human genome assembly version GRCh38.p10. For each trait-associated SNP, we record the effect allele reported by the GWAS catalog for PTS score calculation. SNP effect alleles are all indexed to the positive strand.

Two classes of trait-specific SNP sets were curated for subsequent PTS calculation: (1) individual trait SNP sets parsed directly from the NHGRI-EBI GWAS Catalog annotations, and (2) higher-order trait SNP sets organized according to the EBI Experimental Factor Ontology. For the individual trait SNP sets, text strings from the NHGRI-EBI GWAS Catalog were first normalized in order to import the data into an SQL database. Primary SNP sets were created by querying the database for each unique entry in the "DISEASE/TRAIT" column. Resulting subsets were then filtered to remove interaction terms, duplicate variants, and multi-allelic variants that do not have defined effect alleles. Effect and non-effect alleles were swapped as needed to be consistent within a set (*e.g.*, all effect alleles in the "Type 2 Diabetes" set should increase the risk of diabetes). Finally, trait SNP sets with fewer than three variants were removed after the other filters were applied.

For the higher-order trait SNP sets, the EBI Experimental Factor Ontology was parsed to obtain terminal nodes and subterminal (internal) nodes with high numbers of connected terminal

nodes. First, the terminal node trait descriptors were used to create trait SNP sets from the filtered NHGRI-EBI GWAS Catalog annotations described in the previous paragraph. The majority of the resulting trait SNP sets were identical to the existing individual trait SNP sets taken directly from the NHGRI-EBI GWAS Catalog. The resulting trait SNP sets that differed from the existing sets were retained as additional sets for subsequent PTS calculation. Second, trait descriptors from the highly connected subterminal nodes were used to create higher-order, functionally coherent trait SNP sets from the filtered NHGRI-EBI GWAS Catalog annotations. The rationale for this approach was to maximize the ability to calculate PTS for the traits reported in the NHGRI-EBI GWAS Catalog. For example, there are a number of GWAS studies for which the NHGRI-EBI GWAS Catalog only reports a single SNP for a given trait, and thus do not yield sufficient information for PTS calculation. Combining identical or related traits into more densely populated higher-order SNP sets allows for greater trait coverage of the GWAS Catalog for PTS calculations.

All of the trait descriptors parsed from the Experimental Factor Ontology were hierarchically organized into a custom ontology containing 818 discrete traits, a browsable visualization of which is available online at https://gadget.biosci.gatech.edu/ontology.html. This custom ontology was created to yield a simplified and more intuitive organizational scheme for human phenotypes, which we used to classify our trait SNP sets into 11 functionally related categories for visualizing PTS results: aging, brain health and disorders, cancer, diabetes, general health, heart and health disorders, immune disease and disorders, miscellaneous, obesity, pulmonology, and reproductive health.

2.3.3 Individual and population-specific SNP variant data

The GADGET web server uses publicly available genotype data from the 1000 Genomes Project (1KGP) Phase 3 data release [39] to compute genome-wide PTS for 2,504 individuals from 26 worldwide populations, which are organized into five continental (or super) population groups according to the 1KGP scheme: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). SNP variant genotype data for these individuals were downloaded as VCF files from the 1KGP website at http://www.internationalgenome.org/data. The VCF files were processed to remove SNP variants with >5% missingness, and the remaining variants were annotated with SnpEff [40]. A customized version of the Gemini v0.20.0 application [41] was used to import the resulting filtered and annotated VCF files into an SQLite3 database. The same database was populated with trait and GWAS citation information for all of the trait SNP sets created as previously described. The resulting combined database is queryable by chromosomal position, rsID, gene symbol, trait name, and PMID.

2.3.4 Genome-wide polygenic trait scores

Genome-wide PTS are calculated for individual genome sequences from the processed 1KGP SNP variant data using the curated trait SNP sets described previously (Explore mode) or with user-supplied SNP sets that correspond to traits of interest (Compute mode). In the Explore mode, unweighted PTS ($uPTS$) are calculated as the normalized sums of the number of effect alleles found in the genome for all trait-associated SNPs:

$$uPTS = \sum_{i=1}^{n} EA_i \Big/ \sum_{i=1}^{n} A_i \qquad \textbf{(1)}$$

where $EA_i \in \{0, 1, 2\}$ are homozygous absent, heterozygous, and homozygote present effect alleles for each trait-associated SNP i and $A_i \in \{0, 1, 2\}$ are the total number of alleles with basecalls at each SNP i. PTS are only computed for cases where there are at least 3 SNP positions with basecalls, *i.e.,* when $\sum_{i=1}^{n} A_i \geq 6$, thereby eliminating the possibility of any division by zero error. In the Compute mode, PTS can be computed for user-supplied SNP sets as either unweighted or weighted sums of the number of effect alleles. Weighted PTS (*wPTS*) employ effect size estimates, either odds ratios or β-values, to weight the numbers of observed effect alleles for each trait-associated SNP:

$$wPTS = \sum_{i=1}^{n} (EA_i \times es_i) \Big/ \sum_{i=1}^{n} A_i \qquad \textbf{(2)}$$

where es_i is the SNP-specific effect size estimate.

2.4 Results

2.4.1 Explore mode

In the Explore mode of the GADGET web server, users can visualize the global distributions of genome-wide PTS for 821 polygenic traits organized into 11 phenotypic categories. For each trait, unweighted PTS (equation (1)) are calculated for the 2,504 individual genomes from the 1KGP, and population-specific PTS distributions are shown for 26 global populations organized into 5 continental (super) population groups. The resulting population-specific PTS are visualized as scaled circles on a global map as well as population-specific box plots. The area (A) of the circle for each population (i) is computed as: $A_i = \pi r^2$, where $r = 10 \times (2^{PTS_i/\max PTS})$. The among population variance of trait-specific PTS is measured using

ANOVA, for the five continental population groups, with F-statistics, *P*-values and false discovery rate *q*-values reported in the trait table. A detailed summary of population-specific PTS values, along with the results of the ANOVA analyses, are provided in the 'Summary statistics' field.

Figure 3 shows an example of the Explore mode output for the trait diisocyanate-induced asthma, which shows the most extreme population-specific PTS distributions for any of the pre-computed traits. Diisocyanates are chemical building blocks used to make a wide array of polyurethane products and represent a ubiquitous environmental contaminant. They are a leading cause of workplace respiratory problems and representative of a large class of environmental triggers for respiratory distress [42, 43]. Accordingly, diisocyanate-induced asthma has been investigated by GWAS in an effort to elucidate the genetic architecture of environmentally triggered asthma [44]. Results generated by the GADGET web server show that individuals from African populations have a far higher genetic risk for environmentally triggered asthma than any other population group, as measured by their diisocyanate-induced asthma PTS. The East Asian and Admixed American population groups, which show similar diisocyanate-induced asthma PTS distributions, have the next highest genetic risk profiles for this trait. In contrast, European populations show uniformly low diisocyanate-induced asthma PTS.

These PTS distributions reflect known health disparities for asthma, underscoring the potential utility of comparing PTS across global population groups for investigating the genetic basis of population-specific health outcomes. The results are consistent with previous work showing a relationship between African genetic ancestry and asthma risk in African Americans [45]. Furthermore, in the United States, African-Americans have the highest prevalence of environmentally triggered asthma, followed by Hispanics and East Asians, with European

Americans showing relatively low levels of asthma (https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=15) [46].

Figure 3. Example output for the GADGET Explore mode.

Users can explore global PTS distributions for 821 traits organized into 11 phenotypic categories. The summary table shows traits in descending order of their ANOVA F-statistics, measuring the extent of among population PTS variation, alongside their statistical significance values (P and q). Example results are shown for the highlighted trait diisocyanate-induced asthma. Scaled circles are used to represent population-specific PTS values on a global map. Box-plot PTS distributions are shown for all 26 global populations and the 5 continental (super) population groups. Users have the option to view all the SNPs and effect alleles used to compute PTS for the displayed trait.

2.4.2 Compute mode

In the Compute mode of the GADGET web server, users can supply their own SNP sets in order to analyze global PTS distributions for their traits of interest. The required fields for user-supplied trait SNP tables are: rsIDs, the identity of the effect allele, trait name, and effect size estimates. PTS for the 1KGP individuals and populations can be computed as unweighted or weighted, and users can supply SNP sets for one or more traits of interest in a single file. The SNP set input file format requirements are specified on the website along with an example SNP table that can be downloaded and/or run on the server.

Figure 4 shows the Compute mode output for acute kidney disease based on the example SNP table that is found on the website. These SNPs were curated from a trans-ethnic meta-analysis of five acute kidney disease GWAS, wherein SNP effects were inferred separately for African, European, and Native American ancestry groups [47]. The example input SNP table for this trait considers SNP effects separately for African-American (AfrAm), American Indian (AmInd), and European American (EurAm) GWAS-implicated SNPs, following the convention of the original paper, as can be seen in phenotype column labels. Once the PTS are computed for the three distinct SNP sets, users can toggle among the results for each set using the dropdown menu ('Choose a phenotype to explore'). In this way, the extent to which PTS are influenced by the population ancestry of the study subjects in the GWAS can be assessed. For this particular trait, the global PTS distributions are highly similar across each of the three ancestry-specific SNP sets. African populations consistently show the highest PTS distributions for acute kidney disease, with the European and Admixed American groups being intermediate and the two Asian population groups showing the lowest PTS.

Figure 4. Example output for the GADGET compute mode.

Users can supply their own trait SNP sets for PTS calculation and global PTS distribution visualization. An example trait SNP table, for acute kidney disease, is shown here. This trait is broken down into three phenotypes based on the ancestry-origin of the GWAS SNPs used for PTS calculation. PTS are calculated for all phenotypes, and users can explore each phenotype individually. As with the pre-computed PTS shown in the explore mode, PTS calculated from user-supplied SNP sets are visualized on a global map and as population-specific box plots. ANOVA statistics are shown on the plot for the 5 continental (super) population groups.

2.5 Discussion

2.5.1 Methods for calculating PTS

There are a number of different factors that need to be considered when choosing the specific set of SNPs to be used in PTS calculation for any given trait (BioRxiv: https://www.biorxiv.org/content/early/2017/02/05/106062). The most fundamental decision relates to the number of SNPs to include in a trait set. At the extreme ends of the spectrum, there is the top-SNP approach, whereby only SNPs that reach genome-wide significance are used for either unweighted or weighted score calculation, versus the all-SNP approach, whereby effect sizes are used to weight the phenotypic contributions of all the SNPs that were genotyped in a given study. Between these two extremes, PTS calculation methods can use different GWAS *P*-value thresholds to determine whether SNPs should be included in a trait set. The approach that the GADGET web server uses to calculate PTS can be considered as a soft version of the top-SNP approach since it employs a fairly stringent *P*-value threshold of 10^{-5}, which is nevertheless far more inclusive than the standard GWAS genome-wide significance threshold of 10^{-8}. Our approach is also distinguished by the fact that it sometimes combines SNPs from multiple GWAS into single trait sets. We have found that this approach provides additional resolution for PTS calculation, based in part on the use of larger numbers of SNPs for PTS calculation. Since the effect sizes between multiple studies may not be directly comparable, the pre-computed PTS reported in the server's explore mode are calculated via the unweighted approach. The option for users to supply their own SNP sets provides more flexibility for the computation of PTS, both with respect to the number of SNPs that can be used as well as the weighting scheme.

2.5.2 Genetic ancestry effects on PTS calculation

The vast majority of GWAS have been conducted in populations with European ancestry [16, 33], and the extent to which GWAS-implicated variants replicate across populations remains a matter of contention [32]. On the one hand, a number of trans-ethnic studies have shown that the majority of GWAS implicated variants replicate across populations [48-50]. This is even true for traits such as type 2 diabetes [51, 52], which shows highly population-specific PTS distributions [53, 54]. Furthermore, while the same tag SNPs may not reach genome-wide significance in distinct populations, the haplotypes that they mark are often found to replicate among populations. Nevertheless, even SNPs that replicate between populations could differ with respect to population-specific effect size and explanatory power. Furthermore, a recent study showed that the effects of demographic history on allele frequencies could reduce the accuracy of PTS calculated among divergent populations; for example, PTS for the highly heritable trait height was found to be unreliable across populations [31]. Even GWAS variants that do replicate across populations can show substantial heterogeneity with respect to effect sizes in different populations (BioRxiv: https://www.biorxiv.org/content/early/2017/09/15/188094).

In any case, the results reported by our web server should be interpreted with caution in light of the fact that population-specific PTS will inevitably be generated from SNPs implicated by GWAS on subjects with distinct ancestries. Thus, the PTS distributions that we show may best be considered as hypotheses that can be used to stimulate and guide further investigations. It is also worth noting that, as we illustrated in the example for acute kidney disease, the Compute utility provided on our webserver, whereby users provide their own SNP sets for traits of interest, provides one way to explore whether and how the ancestry of GWAS study subjects influences population-specific distributions of PTS. In addition, the comparison of unweighted and weighted

scores for user-supplied SNP sets can be used to evaluate the effect of ancestry-specific effect size estimates on PTS population differences.

2.6 **Conclusion**

2.6.1 Interpreting PTS differences across populations

As mentioned previously, the meaning of PTS differences across human populations very much remains an area of active investigation, and there are numerous possible interpretations for such results. It is important to keep these alternative explanations in mind when interpreting the worldwide PTS distributions generated by the GADGET server. Some of the possible explanations for PTS differences among global populations are: (1) the genetic predisposition to the trait differs among populations, (2) the top SNPs used for the analysis differ among populations, but the overall genetic predisposition for the trait would balance out if additional SNPs were included in the PTS calculation, (3) the apparent population differences in the genetic predisposition for any given trait could disappear due to heterogeneous effect sizes among populations, (4) observed population differences in PTS could be due to stochastic effects related to demographic factors (*e.g.,* genetic drift). These are just some of the possible explanations; the list is by no means exhaustive. In addition, problems with the original GWAS studies or issues with the accuracy of the GWAS database used to generate trait-associated SNP sets could also cause problems with global PTS distributions. In light of these caveats, PTS results generated by GADGET should be treated with caution.

Finally, users are cautioned not to use GADGET to draw conclusions regarding the genetic basis of racial differences. GADGET allows for the interrogation of PTS differences across human population groups characterized as part of the 1KGP, which are defined by geographic origin and

distinguished by genetic ancestry. We make no attempt to delineate racial groups from these populations and the extent to which racial classifications accurately reflect genetic ancestry remains a matter of contention [55-57].

CHAPTER 3. THE PHENOTYPIC CONSEQUENCES OF GENETIC DIVERGENCE BETWEEN ADMIXED LATIN AMERICAN POPULATIONS: ANTIOQUIA AND CHOCÓ, COLOMBIA

3.1 Abstract

Genome-wide association studies have uncovered thousands of genetic variants that are associated with a wide variety of human traits. Knowledge of how trait-associated variants are distributed within and between populations can provide insight into the genetic basis of group-specific phenotypic differences, particularly for health-related traits. We analyzed the genetic divergence levels for (i) individual trait-associated variants and (ii) collections of variants that function together to encode polygenic traits, between two neighboring populations in Colombia that have distinct demographic profiles: Antioquia (*Mestizo*) and Chocó (Afro-Colombian). Genetic ancestry analysis showed 62% European, 32% Native American, and 6% African ancestry for Antioquia compared to 76% African, 10% European, and 14% Native American ancestry for Chocó, consistent with demography and previous results. Ancestry differences can confound cross-population comparison of polygenic risk scores (PRS); however, we did not find any systematic bias in PRS distributions for the two populations studied here, and population-specific differences in PRS were, for the most part, small and symmetrically distributed around zero. Both genetic differentiation at individual trait-associated SNPs and population-specific PRS differences between Antioquia and Chocó largely reflected anthropometric phenotypic differences that can be readily observed between the populations along with reported disease prevalence differences. Cases where population-specific differences in genetic risk did not align with observed trait (disease) prevalence point to the importance of environmental contributions to the phenotypic

variance for both infectious and complex common diseases. The results reported here are distributed via a web-based platform for searching trait-associated variants and PRS divergence levels at http://map.chocogen.com.

3.2 Background

The genetic basis of human phenotypic diversity is both an issue of fundamental evolutionary interest and critical to a deeper understanding of health disparities. Early genetic linkage analyses, and more recent GWAS, have uncovered thousands of genetic variants that are associated with a wide variety of human traits [34, 58]. Investigations of how trait-associated genetic variants are distributed within and between populations have the potential to shed light on the genetic architecture of human phenotypic diversity, particularly as related to disease prevalence disparities [59, 60].

The power of this approach has long been apparent for single locus traits. Population-specific distributions of rare and highly penetrant variants that cause Mendelian diseases are responsible for a wide variety of population health disparities, such as sickle-cell anemia (OMIM: 603903), cystic fibrosis (OMIM: 219700) and Tay-Sachs disease (OMIM: 272800). Of course, the vast majority of human traits are encoded by multiple loci, each of which contributes only a small fraction of the total trait variance [61]. Individuals' genomic predispositions to such multi-locus traits can be captured by PRS – also known as polygenic risk scores, genome-wide risk scores, or genetic risk scores – which are calculated as (weighted) sums of the total number of trait-associated or trait-increasing alleles present in the genome [35, 62]. Changes in PRS distributions across populations have been taken as evidence of polygenic selection on a number of anthropometric [63-65], neurological [66], and disease-related traits [67].

Despite their apparent potential for discovering genetic changes that underlie phenotypic divergence among populations, recent studies have underscored a number of challenges entailed by cross-population comparisons of PRS. Systematic differences in allele frequencies, proportions of ancestral versus derived alleles, and patterns of linkage disequilibrium can yield large shifts in PRS distributions that do not necessarily reflect observed phenotypic differences among populations [23, 30, 31]. Furthermore, the fact that the vast majority of GWAS have been conducted on cohorts of European ancestry [16, 32, 33] yields PRS that are far more accurate for European populations compared to other, less-studied global population groups [24]. In light of these challenges, the goals of this study were to: (1) characterize the genetic ancestry patterns for diverse populations from within a single Latin American country, (2) evaluate the impact of ancestry differences between these populations on the genetic variants associated with anthropometric and disease traits, and (3) consider observed differences in the frequencies of trait-associated variants in light of known phenotypic differences between the populations.

Recently admixed populations hold great promise for studies aimed at characterizing the genetic basis of phenotypic divergence [68], but studies of cross-population PRS have yet to focus explicitly on admixed populations. Furthermore, studies of this kind have not focused on diverse populations that often co-exist in close physical proximity in the modern world. Our research group is focused on the study of admixed American populations, with the broad aim of relating differences in ancestry to genetic determinants of health-related phenotypes [69-74]. Latin American populations are particularly interesting for studies of this kind, given their high levels of genetic admixture among ancestral African, European, and Native American population groups [75-78]. Populations within and between Latin American countries are characterized by different levels of continental and regional ancestry. We have been studying two neighboring populations

from Colombia – Antioquia and Chocó – that are distinguished by a combination of close proximity and divergent demographic profiles. We previously found that sample donors from Antioquia show primarily European genetic ancestry, whereas donors from Chocó show majority African ancestry [79, 80], and we showed that this divergent genetic ancestry, and the allele frequency differences that it entails, lead to an increase in the predicted risk of type 2 diabetes (T2D) in Chocó compared to Antioquia [54]. T2D is an intensively studied disease, and this pattern of greater predicted T2D risk in Chocó holds irrespective of the ancestry of the GWAS cohorts used for risk allele discovery [81]. For this study, we performed a broader survey of the genetic divergence levels for trait-associated variants and differences in PRS for these two admixed Colombian populations, and we considered the results of these comparisons in light of known (observable) demographic and phenotypic characteristics for these two populations.

3.3 Materials and Methods

3.3.1 Genomic data

The sources of genomic data used for this study are shown in Table 1. Whole genome genotype data for the population of Chocó, Colombia, were taken from the ChocoGen research project https://www.chocogen.com [79, 80]. The ChocoGen project was conducted with the approval of the Ethics Committee of the Universidad Tecnológica del Chocó (ACTA Nº 01-v1) following the Helsinki ethical principles for medical research involving human subjects. All sample donors signed informed consent documents. Whole genome sequence data for the population of Antioquia, Colombia were taken from the phase 3 data release of the 1000 Genomes Project [82]. The 1000 Genomes Project human genome sequence data are de-identified and made publicly available for research use without restriction.

The whole genome sequence and genotype data for continental reference populations from Africa, the Americas, and Europe were taken from the 1000 Genomes Project and a collection of previously characterized Native American populations [83]. The Native American genotype data are de-identified and made publicly available for research according to the terms of a data use agreement from the Universidad de Antioquia. A list of all bioinformatics programs and databases used for the analyses is shown in Table 2.

3.3.2 Genetic ancestry analysis

Whole genome genotype and sequence variant data were merged using PLINK version 1.9 [84], with SNPs common to all three data sources retained for subsequent analysis and SNP strand orientations corrected as needed. The merged SNP set was phased using ShapeIT version 2.r837 with the 1000 Genomes Project haplotype reference panel [85, 86], and PLINK was used to prune linked SNPs from the phased genotype dataset with an r^2 threshold of 0.1. The merged and pruned SNP set was used to infer three-way continental ancestry ($f_{African}$, $f_{European}$, $f_{NativeAmerican}$) for Antioquia and Chocó using the program ADMIXTURE version 1.3.0 [87] run in unsupervised mode, with K=3 continental ancestral groups corresponding to the African, European, and Native American reference populations shown in Table 1. SNP allele frequency differences and Fixation Index (F_{ST}) values between Antioquia and Chocó were computed from the merged SNP set using PLINK. F_{ST} values were calculated using the Weir and Cockerham estimator [88]. Ternary plots were constructed using the inferred global ancestry fractions for each individual and the position of each individual (point) within the triangle is a composition of the individual's three ancestry components:

$$\frac{1}{2}\cdot\frac{2A+N}{E+A+N},\frac{\sqrt{3}}{2}\cdot\frac{N}{E+A+N} \quad \textbf{(3)}$$

where E, A, and N are the inferred European, African, and Native American ancestry components.

3.3.3 SNP trait-associations and polygenic scores

SNP trait-associations were taken from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) [89], with the SNP rsid number, effect allele, effect size, and study population recorded for all associations. Effect alleles are operationally defined as the allele for any given SNP that is associated with cases, for case-control GWAS, or with an increase in the trait under consideration for quantitative trait GWAS. The SNP associations used here are limited to biallelic variants, do not include SNP interactions, and are all significant at $P<1\text{x}10^{-5}$ (# of SNPs = 107,784). SNP associations were grouped into polygenic traits using the NHGRI-EBI GWAS Catalog trait terms (# of traits = 2,382), which are derived from the EBI Experimental Factor Ontology (https://www.ebi.ac.uk/efo/) [90]. After filtering, 65,283 (60.5%) SNPs remained. Of the 27,886 (39.5%) associations excluded: 25,305 (23.5%) had an unknown or unreported effect allele (effect allele = "?"); 14,615 (13.5%) had multiple reported effect alleles for the same trait and reported effect alleles were not strand-flips (i.e., A and C); and 2,581 (2.4%) had no associated rsID (i.e., the variant is given by chromosomal location, chr1:2345).

Whole genome genotype data from Chocó were imputed up to 1000 Genomes phase 3 variant calls using the program IMPUTE2 version 2.3.2 [91, 92] and the 1000 Genomes Project haplotype reference panel. Imputed sites were retained for subsequent analysis if they had a 95% imputation rate across samples and an INFO score > 0.4. The imputed data from Chocó were merged with the whole genome sequence variant data from Antioquia using PLINK.

Polygenic risk scores (PRS) were computed for each GWAS trait i as the sum of the effect alleles across all trait-associated SNPs as previously described (Equation (1): The "top-SNP" approach, i.e., the use of only highly significantly associated SNPs for PRS calculation, was chosen to mitigate confounding effects of population structure on PRS comparisons. Unweighted PRS were used to allow for combining SNP trait-associations across multiple studies, each with distinct effect size estimates. We opted not to use linkage disequilibrium (LD) pruning for PRS calculation to facilitate direct comparison of PRS between populations with divergent LD structure.

For each of the three continental ancestry components ($f_{African}$, $f_{European}$, $f_{NativeAmerican}$), individuals' continental ancestry fractions were regressed against their PRS using unweighted ordinary least squares regression (OLS):

$$PRS_i = \alpha + \beta x_i + \varepsilon_i \quad \textbf{(4)}$$

where $PRS_i$is the predicted polygenic risk score for individual i; α and β are constants describing the intercept and slope, respectively; x_i is the ancestry fraction for individual i; and ε_i is an error term describing the deviation from the fitted line. The resulting OLS produces: β_0, the model β or slope; the standard error of the model; the r^2 value describing the model's fit; the model t-statistic; and a two-tailed P-value.

Trait-associated SNPs were mapped to the nearest genes for pathway enrichment analysis using the ENSEMBL rsID to HGNC mapping API (getBM) provided as part of the biomaRt R package (attributes = refsnp_id, ensemble_gene_stable_id, hgnc_symbol, entrezgene_id; filter = snp_filter & ensembl_gene_id; values = GWAS Catalog SNP rsIDs). SNPs that did not return an

HGNC mapping were discarded. Genes were assigned population-specific effect allele frequency difference values ($\Delta f = f(EA_{Ant}) - f(EA_{Cho})$) based on the SNP with the maximum effect allele frequency difference: $\max|\Delta f_{g,i}|$, where g is a trait-associated gene, and i is ith SNP in the gene g. The Δf values for all mapped trait-associated genes were used to create population-specific gene lists for pathway over-representation analysis using the hypergeometric test implemented in the "enricher" function from the clusterProflier version 3.14.0 R package [93]. Briefly, for each gene, the sign on Δf was used to assign a gene to the Antioquia (positive) or Chocó (negative) gene lists. For each population-specific gene list and for each gene set, a hypergeometric test was performed using: $\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$, where m is the number of population-specific genes, k is the number of population-specific genes in the gene set, n is the number of genes in gene set, and N is number of genes in the background. Gene sets from the KEGG, MSigDB (http://software.broadinstitute.org/gsea/msigdb/), and PID (http://pid.nci.nih.gov) were used in the enrichment analysis.

The relative predicted disease risk and observed disease prevalence for Antioquia and Chocó were computed as the $\log_2$ odds ratio for the effect of allele frequencies and the reported age-adjusted disease prevalence values for Chocó/Antioquia. For each disease-associated SNP, its log odds ratio is computed as: $\log_2 \frac{p_{Cho}/q_{Cho}}{p_{Ant}/q_{Ant}}$, where p_{pop} is the population-specific frequency of the effect allele and q_{pop} is the population-specific frequency of the non-effect allele. The log odds ratio values for all associated SNPs were summed for each disease. The log odds ratio for disease prevalence is computed as: $\log_2 \frac{Disease_{Cho}/No\ disease_{Cho}}{Disease_{Ant}/No\ disease_{Ant}}$. Disease prevalence ($Disease_{pop}$ and $No\ disease_{pop}$) was defined as the population- and age-adjusted prevalence per 100,000 and $(100{,}000 - prevalence)$ reported for each department in 2017 and were taken from Colombian

governmental and non-governmental resources (see *Demographic, lifestyle and disease prevalence data* section below).

3.3.4 Demographic, lifestyle and disease prevalence data

A variety of sources was used to curate demographic, lifestyle, and disease prevalence data for Antioquia and Chocó. The 2005 general census published by the Colombian Departamento Administrativo Nacional de Estadística (DANE) was used for demographic and socio-economic status data [94]. Disease prevalence data were taken from three epidemiological databases: (1) Cuenta de Alto Costo (https://cuentadealtocosto.org/), (2) Observatorio de Diabetes de Colombia (http://www.odc.org.co/), and (3) the Sistema Integral de Información de la Protección Social (https://www.minsalud.gov.co/salud/Paginas/SistemaIntegraldeInformaciónSISPRO.aspx). Diet and lifestyle data were taken from the Colombian national nutritional survey [95].

3.4 Results and discussion

3.4.1 Demography and genetic ancestry in Antioquia and Chocó

Antioquia and Chocó are Colombian administrative departments (i.e., states) that are located in the northwestern part of the country and share a common border (Figure 5A). Chocó runs along the Pacific coast and borders Panamá to the north; it is the only department in Colombia with Pacific and Atlantic coasts. Antioquia is situated due east of Chocó, in the interior of the country, and also has a short Atlantic coastline. Despite their close proximity, the two departments have very distinct geography and climate, as well as distinct historical and demographic profiles. Antioquia occupies the mountainous Andean region of the country and is traversed by the Western and Central Andes mountain ranges. According to the 2005 census, approximately 89% of the

Antioquia population identifies as white or mestizo compared to 11% black or Afro-Colombian and less than 1% Indigenous. Chocó lies along the lowland Pacific coastal region and is almost entirely covered by dense tropical rainforest. The climate is hot and humid, and the region receives some of the highest rainfall totals in the world. The population of Chocó identifies as 8% Afro-Colombian, 13% Indigenous, and 5% white or mestizo.

Genome-wide variant data from Antioquia and Chocó were compared to data from African, European, and Native American continental reference populations to infer the patterns of genetic ancestry and admixture in the two Colombian populations. The genetic ancestry of Antioquia and Chocó reflect their distinct historical founding populations, physical and cultural barriers to migration, and current demographic profiles (Figure 5B and C). Antioquia shows predominantly European genetic ancestry (average ± standard error; 62% ± 1.55) followed by Native American (32% ± 1.24) and then African (6% ± 0.83) components; whereas, Chocó has primarily African genetic ancestry (76% ± 1.65) with approximately equal parts Native American (14% ± 0.83) and European (10% ± 1.03) ancestry.

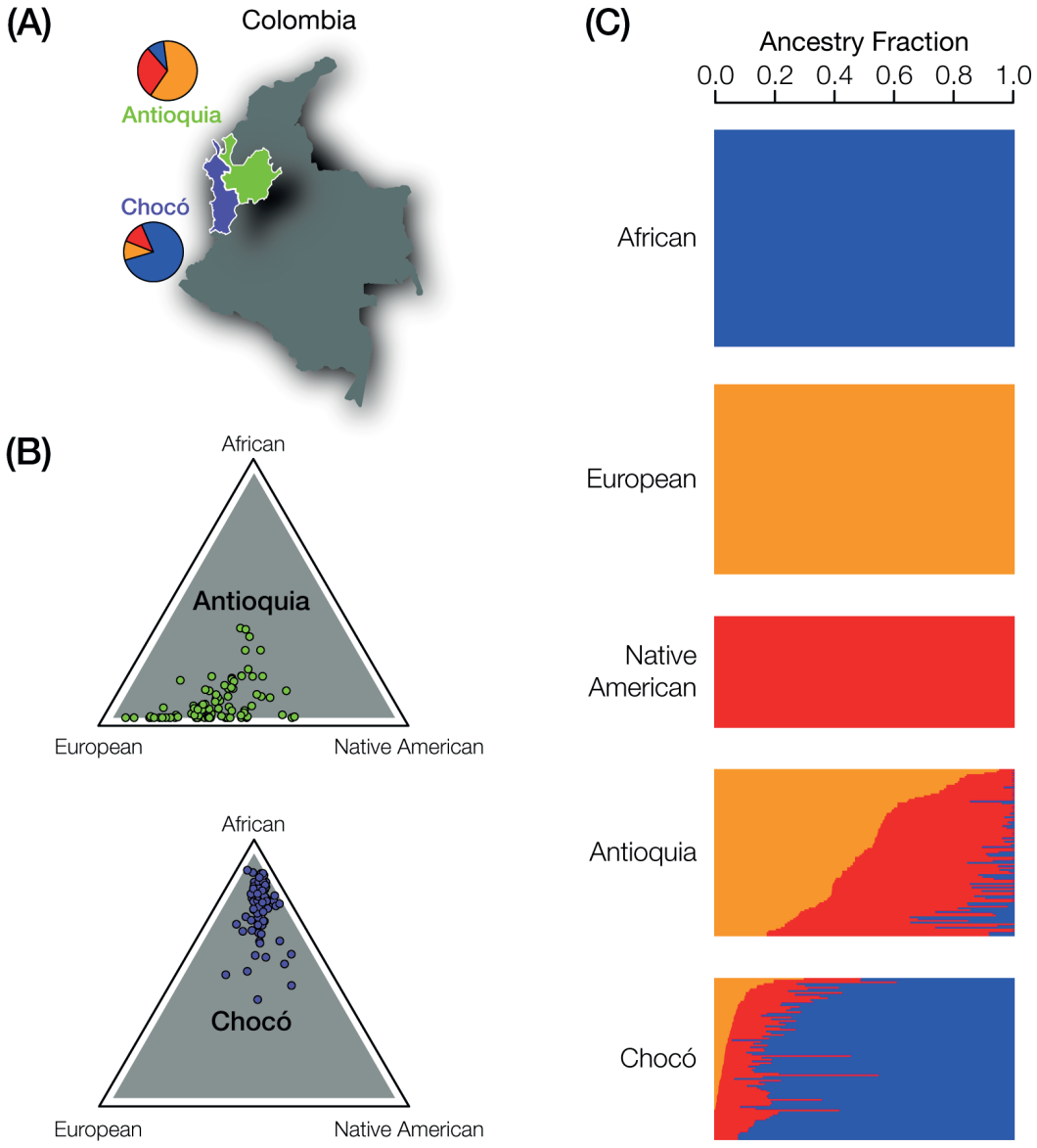

Figure 5. Genetic ancestry in Antioquia and Chocó.

(A) The locations of the Colombian administrative departments of Chocó (purple) and Antioquia (green) are shown along with pie charts indicating the average continental ancestry fractions: African (blue), European (orange), and Native American (red). (B) Ternary plots showing the relative contributions of African, European, and Native American ancestry to individuals from Antioquia (green) and Chocó (purple). (C) ADMIXTURE plot showing the continental ancestry fractions for African (blue), European (orange), and Native American (red) reference populations together with Antioquia and Chocó.

3.4.2 Single variant divergence and phenotypic associations

The potential impact of ancestry differences between Antioquia and Chocó on the genetic architecture of phenotype and function was assessed for individual SNP trait-associations (Figure 6). A total of 47,398 SNP trait-associations were curated and evaluated with respect to the extent and direction of differentiation between Antioquia and Chocó. Population differentiation was measured by effect allele F_{ST} values and frequency differences between the two populations (Figure 6A and B, Table 5). The top 20 most extreme values correspond to both known phenotype and disease prevalence differences between the two populations as well as novel differences (Figure 11). Pigmentation associated variants for both skin and hair show expected differences with lighter skin and hair effect alleles found in higher frequency in Antioquia compared to Chocó. Antioquia also shows higher frequencies of Crohn's and inflammatory bowel disease SNP effect alleles than Chocó. In contrast, Chocó shows higher frequencies of variants associated with prostate and breast cancer along with Alzheimer's and asthma, consistent with known health disparities around the world. Chocó also showed a substantially higher frequency of variants linked to resistance to the malaria parasite *Plasmodium vivax*. Unexpected results include the higher frequency of nicotine use associated SNP effect alleles in Chocó, as tobacco use is known to be lower in Chocó compared to Antioquia, the greater waist-hip ratio in Antioquia, and the increased longevity in Chocó.

Word clouds provide a visual sense of the overall between-population divergence for all trait-associated SNPs, with the most enriched traits highlighted for each population (Figure 6C). The word clouds were generated using all trait-associated SNPs that showed $F_{ST} > 0.2$, 61 SNPs for Antioquia and 98 for Chocó, and therefore provide additional resolution on the divergence of single variant associations between populations. For example, schizophrenia appears in the word

clouds for both populations (Figure 6C), with more weight in Chocó. However, it was not present in the top 20 divergent associations shown in Figure 6 panels A and B. Obesity-related traits appears as overrepresented in Chocó in the word cloud (Figure 6C), even though the most diverged body mass index SNP shows higher frequency in Antioquia (Figure 6A & B). This is due to a preponderance of obesity-associated SNPs among the complete set of variants with $F_{ST} > 0.2$ and is consistent with what is seen via polygenic trait divergence analysis (see next section and Figure 7). Overall, the population divergence observed for single variant associations is consistent with reported health disparities and demographic data in Colombia and Latin American (Table 6).

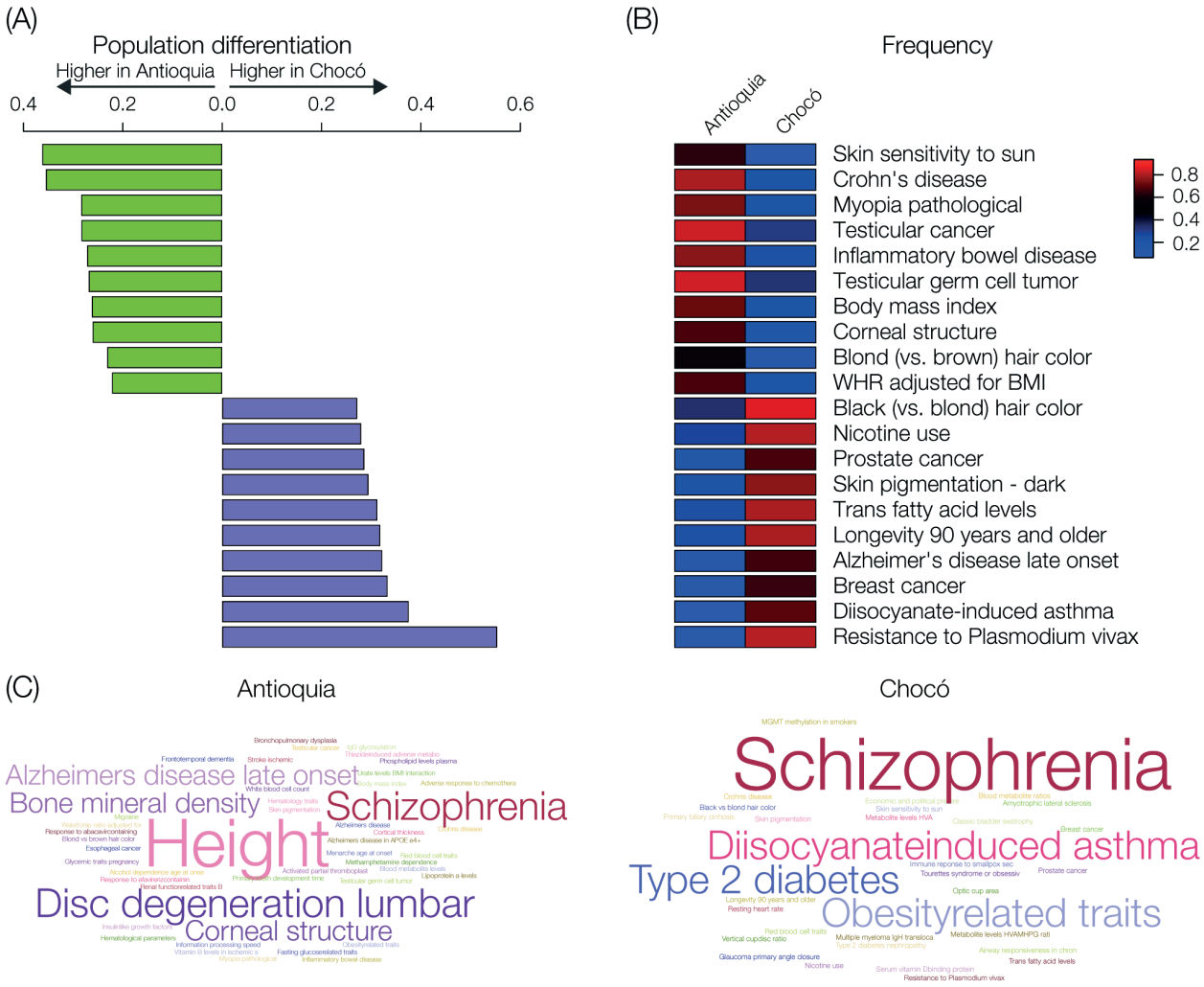

Figure 6. Single nucleotide variant phenotype associations.

(A) Polarized fixation index (F_{ST}) values for divergent trait-associated SNP effect alleles: higher effect allele frequency in Antioquia (left, green) and higher effect allele frequency Chocó (right, purple). The corresponding SNP associations are shown in panel B. (B) Heatmap of population-specific effect allele frequencies (see key) and their SNP associations. (C) Word clouds showing the enrichment of SNP-associated traits for each population.

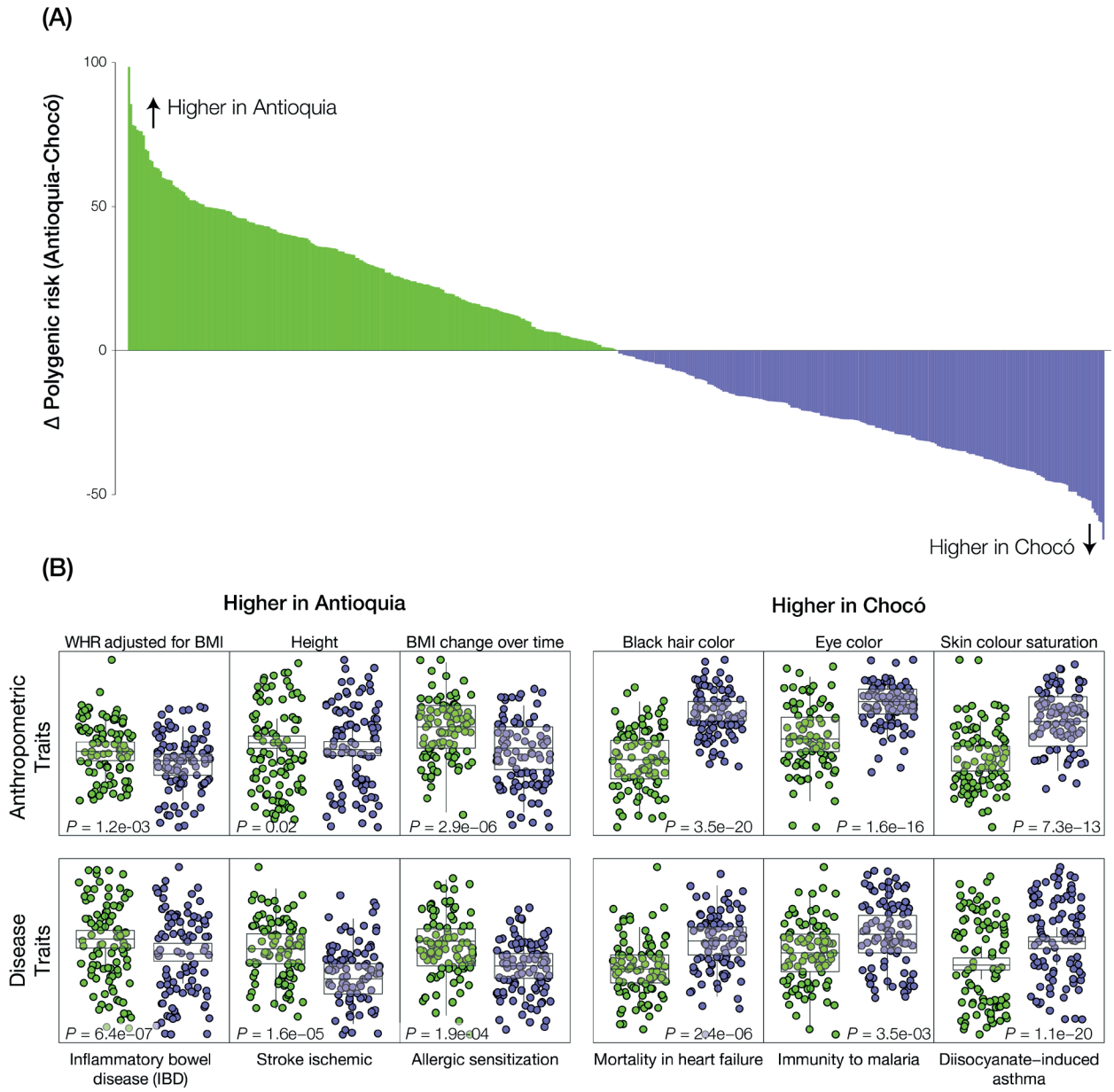

Figure 7. Polygenic risk divergence.

(A) Distribution of the differences in population-average polygenic trait scores (PRS) are shown for significantly divergent traits: higher in Antioquia (above, green) and higher in Chocó (below, purple). (B) Population-specific PRS distributions for examples of anthropometric and disease traits are shown for Antioquia (green) and Chocó (purple) along with the significance levels for the distribution differences. Traits with increased prevalence/risk in Antioquia are shown on the left, with increased prevalence/risk in Chocó are shown on the right.

3.4.3 Polygenic trait divergence

Most human phenotypes are encoded by multiple loci across the genome, each of which contributes to a small fraction of the overall trait variance, i.e., they are polygenic. The relationship between genetic ancestry and polygenic trait architecture in Antioquia and Chocó was assessed by comparing distributions of PRS between the two populations (Figure 7,

Table 7). A total of 1,983 PRS were compared between the two populations, and the overall distribution of $\Delta PRS\ (Ant - Choc)$ is symmetrically distributed around -0.01 (Figure 12), indicating that the differences in genetic ancestry between the populations are slightly biased towards increased predicted risk in Chocó in cross-population PRS inference ($p < 0.001$). This is consistent with theoretical results showing that the divergence of neutral polygenic traits between populations is expected to be small, no different from the expectation for single gene traits, and symmetrically distributed around zero [96, 97]. $\Delta PRS\ (Ant - Choc)$ values for traits that show significantly different mean PRS (Holm-Bonferroni corrected $P<0.05$) are shown in Figure 7A (column D in

Table 7), and population-specific PRS distributions for individual traits of interest are shown in Figure 7B. The specific traits of interest were chosen based on their highly divergent PRS values and their relevance to Colombia owing to the reported public health burden in the country and as reflected by their descriptions in epidemiological and/or census databases.

The individual PRS distributions shown in Figure 7B are organized into anthropometric and disease traits, most of which correspond to the top SNPs from Figure 6. For anthropometric

traits, Antioquia has a higher predicted height and body mass index (BMI), whereas Chocó has higher predicted values for several pigmentation related traits: hair, eye, and skin color. For disease traits, Antioquia has a greater predicted risk for inflammatory bowel disease, ischemic stroke, and allergic sensitization, whereas Chocó has a higher predicted risk for mortality in heart failure, immunity to malaria, and environmentally (diisocyanate) induced asthma. We also explored the impact of GWAS discovery and replication population ancestry on PRS differences for four selected traits from Figure 6 and Figure 7 for which multiple GWAS utilizing different ancestry populations were available: asthma, ischemic stroke, myopia, and type 2 diabetes (Figure 13, Table 8). In all cases, significant differences in predicted population risk profiles were robust to discovery population ancestry, suggesting a shared genetic architecture of risk. In addition, predicted population-specific disease risk profiles are consistent with what has been observed in Colombia (Table 6) as well as with known ancestry-disease associations worldwide: e.g., asthma [98, 99], heart failure [100, 101], irritable bowel disease [102, 103], malaria [104-106], and stroke [107].

We also explored population-specific differences in endophenotypes, with respect to specific pathways and biochemical functions that underlie the observed trait differences, using pathway enrichment analysis (Figure 8). Antioquia shows enrichment for integrin pathways implicated in a number of cancers and inflammatory bowel disease. Chocó shows enrichment for a number of cancer-related pathways, including prostate cancer, which is known to be more prevalent in men of African ancestry [108, 109], as well as T2D and related glycerolipid metabolism pathways.

Given the differences in genetic ancestry seen for Antioquia and Chocó (Figure 5), we evaluated the relationship between individuals' continental genetic ancestry fractions and their

PRS for each trait considered here. It should be noted that, despite the clear differences in the overall ancestry differences seen for the two Colombian populations, almost all individuals analyzed here show substantial admixture with varying fractions of African, European, and Native American ancestry. This fact allowed us to correlate genetic ancestry and PRS along a continuum of continental ancestry fractions (Figure 9). There are significant differences in the magnitude of the PRS correlations among the three ancestry components ($F=4.79$, $P=8.3\times10^{-3}$); African ancestry shows the highest overall correlation with the PRS values of all traits analyzed here, as shown by the median of the distribution, followed by the European and then the Native American ancestry components (Figure 9A). All three populations show a number of apparent cases of high correlations between ancestry and PRS. All traits that show $r^2>0.4$ for any of the three ancestry components are shown in Figure 9B, and individual examples of ancestry × PRS regressions are shown in Figure 9C. Breast cancer PRS is positively associated with European ancestry and negatively associated with African ancestry (Figure 9C), in contrast to what was seen for an individual breast cancer associated variant found at higher frequency in Chocó (Figure 6B). This difference is best explained by the analysis of individual SNPs shown in Figure 6 and the PRS based on multiple SNPs, which are likely to be more reliable, shown in Figure 7 and Figure 9. All ancestry × PRS r^2 values are shown in Table 9.

The high correlations observed between ancestry and PRS could be attributed to artifacts related to uneven cohort sampling in GWAS, as previously discussed, or they could represent actual ancestry-related phenotypic differences between the two populations. The small overall systematic bias in PRS for the two populations (Figure 12), considered together with the fact that most of these ancestry-associations conform to observable anthropometric features and/or previously suggest that these associations reflect real phenotypic differences. However, definitive

proof for this would require individual-level phenotype data, as opposed to the population-level data used here, as well as the use of trait-associated variants that replicate across ancestry-specific GWAS. It should also be noted that these regressions could be confounded by a number of other variables including sex, age, and socioeconomic status that are not available for this study, and which would need to be simultaneously modeled to ensure that the correlations between ancestry and PRS observed here are robust.

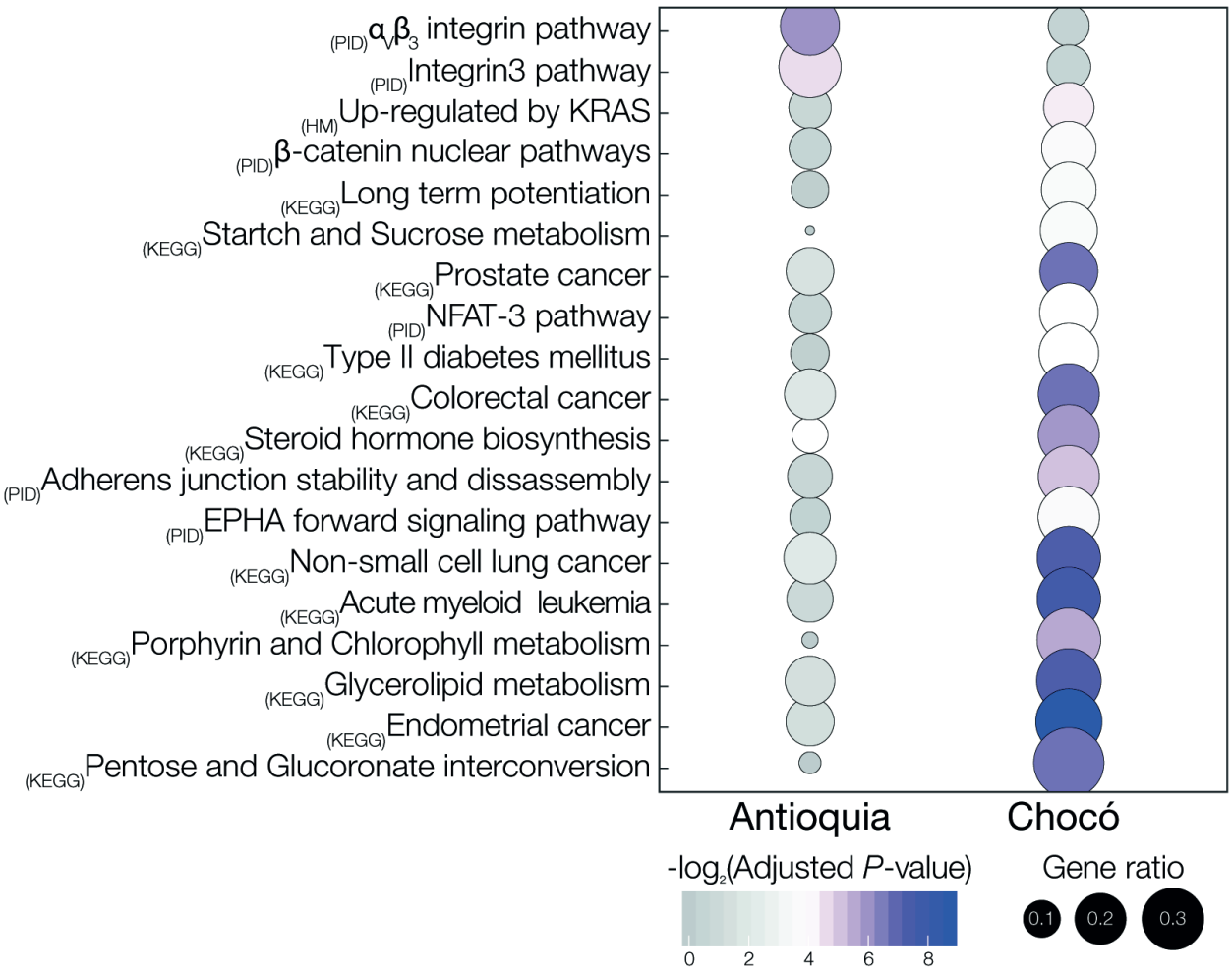

Figure 8. Population-specific differences in trait endophenotypes: pathways and biochemical functions.

Gene set enrichment was used to uncover pathways and functional gene sets that are enriched for divergent associated SNPs in each population. For each pathway or function, circles are scaled to the relative number of implicated genes for each population and colored according to the population-specific levels of enrichment.

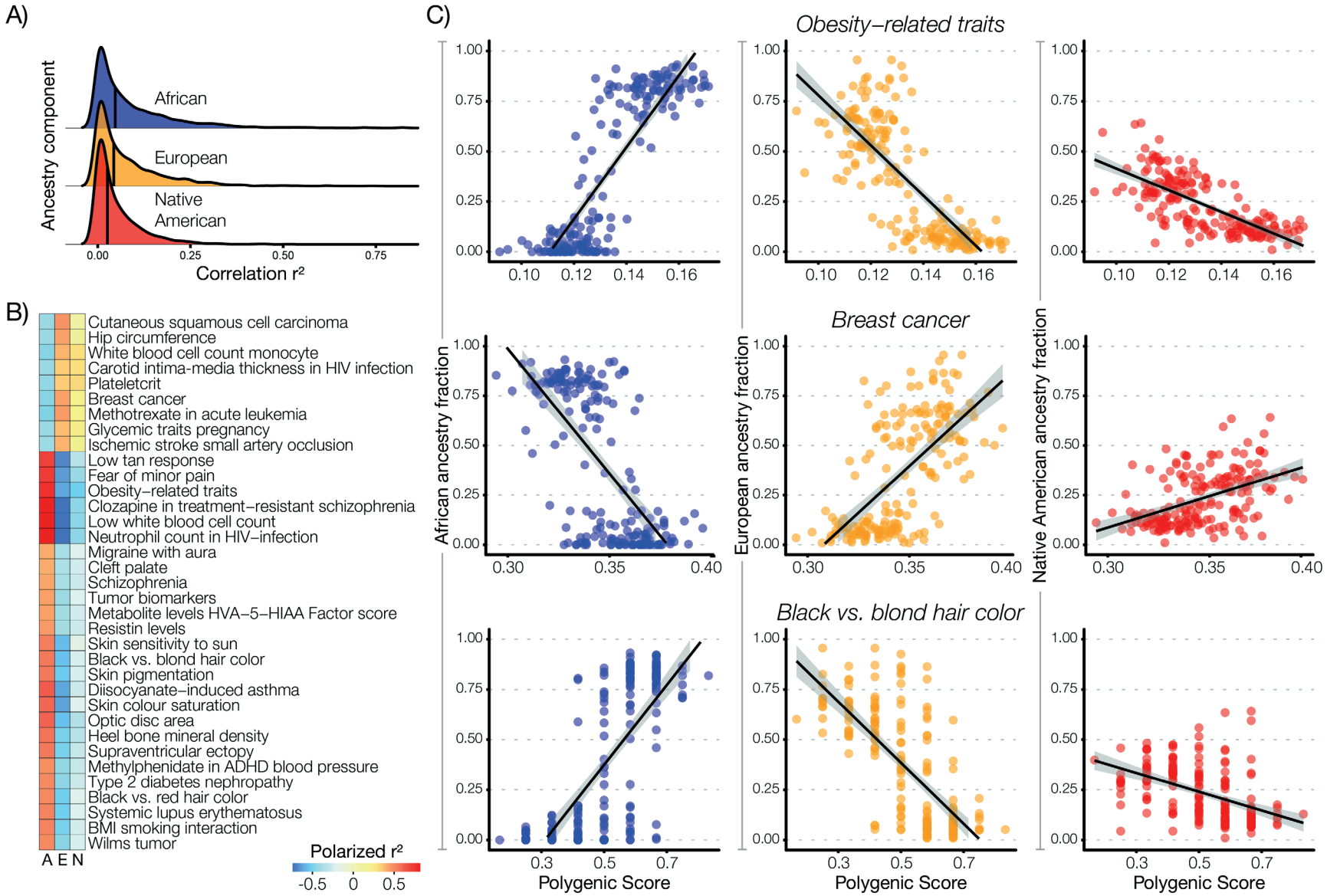
A)
Ancestry component
African
European
Native American
0.00
0.25
0.50
0.75
Correlation r²
B)
Cutaneous squamous cell carcinoma
Hip circumference
White blood cell count monocyte
Carotid intima-media thickness in HIV infection
Plateletcrit
Breast cancer
Methotrexate in acute leukemia
Glycemic traits pregnancy
Ischemic stroke small artery occlusion
Low tan response
Fear of minor pain
Obesity–related traits
Clozapine in treatment–resistant schizophrenia
Low white blood cell count
Neutrophil count in HIV–infection
Migraine with aura
Cleft palate
Schizophrenia
Tumor biomarkers
Metabolite levels HVA–5–HIAA Factor score
Resistin levels
Skin sensitivity to sun
Black vs. blond hair color
Skin pigmentation
Diisocyanate–induced asthma
Skin colour saturation
Optic disc area
Heel bone mineral density
Supraventricular ectopy
Methylphenidate in ADHD blood pressure
Type 2 diabetes nephropathy
Black vs. red hair color
Systemic lupus erythematosus
BMI smoking interaction
Wilms tumor
A E N
Polarized r²
-0.5 0 0.5
C)
Obesity–related traits
Breast cancer
Black vs. blond hair color
African ancestry fraction
European ancestry fraction
Native American ancestry fraction
Polygenic Score
Polygenic Score
Polygenic Score

Figure 9. Genetic ancestry and polygenic trait divergence.

(A) Distributions of the correlations (r^2) between individuals' genetic ancestry fractions– African (blue), European (orange), Native American (red) – and their polygenic trait scores (PRS) for all traits analyzed here. Vertical lines show the median for each distribution. (B) Ancestry x PRS correlations (r^2) polarized by the direction of the correlation (positive or negative) are shown for all traits where $r^2>0.4$ for at least one ancestry component – African (A), European (E), and Native American (N). (C) Examples of polygenic traits with high correlations between ancestry and PRS are shown. Ancestry components are color-coded as in panel A, and for each scatter plot, ancestry fractions (y-axis) are regressed against PRS (x-axis). Linear trend lines with 95% confidence intervals are shown for each regression

.

3.4.4 Predicted versus observed disease risk profiles

Population-specific differences for trait-associated variants, both for single SNP associations and polygenic traits, showed an overall concordance between genetic risk predictions and observed anthropometric and epidemiological profiles for Antioquia and Chocó (Figure 6 and Figure 7). We quantified the relationship between predicted disease risk and observed prevalence for twelve high impact diseases that have been prioritized by the Colombian Ministry of Health via the 'Cuenta de Alto Costo' (http://www.cuentadealtocosto.org/). This analysis was done for complex common diseases, cancers, and infectious diseases (Figure 10). T2D shows the largest difference between predicted disease risk versus observed disease prevalence for Antioquia and Chocó. We previously showed that this difference could be attributed to the higher genetic risk associated with African genetic ancestry and T2D protective environmental factors associated with socioeconomic status in Chocó [54]. In Colombia, environmental factors associated with differences in development across the country appear to have a high impact on the risk of complex common diseases like T2D. A similar, albeit not nearly as extreme, difference can be seen for chronic kidney disease; Chocó has a higher predicted genetic risk but lower prevalence compared to Antioquia. Higher risk for chronic kidney disease has been observed for Afro-descendant populations in other countries [110, 111], consistent with the higher genetic risk for Chocó seen here; thus it may be the case that similar environmental protective factors, with respect to diet and lifestyle, serve as a protective factor for chronic kidney disease in Chocó. Finally, there are large differences in predicted risk (susceptibility) versus observed prevalence for malaria caused by both *Plasmodium vivax* and *P. falciparum*. The population of Chocó has a lower predicted risk for malaria infections, consistent with previous studies on Afro-descendant populations [104-106].

However, both *P. vivax* and *P. falciparum* are far more prevalent in Chocó compared to Antioquia [112-114], thereby explaining the higher malaria prevalence in Chocó.

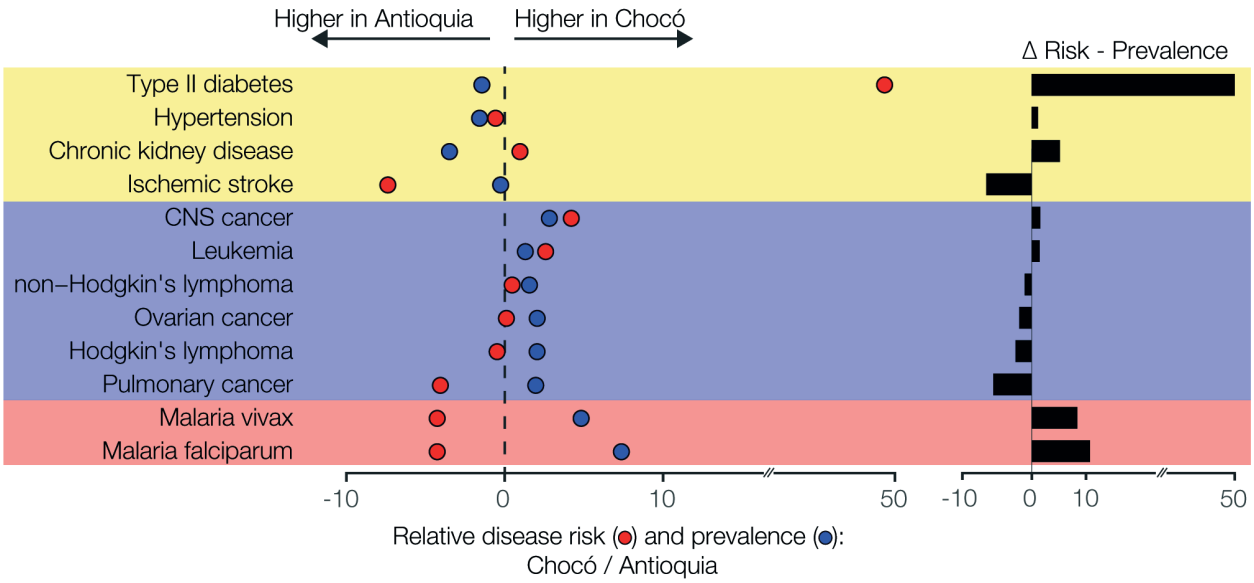

Figure 10. Predicted versus observed disease risk.

Left: For each disease, the predicted genetic risk difference for Antioquia compared to Chocó (red circles) is compared to the observed prevalence of the disease (blue circles). Right: The differences between predicted disease risk minus observed prevalence. Diseases are grouped into bands as complex, common diseases (yellow), cancer (blue), and infectious disease (red). The x-axis values are log odds ratios for population-specific disease risk allele frequencies and observed disease prevalence values, as described in the Materials and Methods.

3.5 Conclusions

Results on the population divergence of trait-associated variants reported here should be interpreted with caution in light of the previously discussed challenges to cross-population genetic risk inference [23, 24, 30, 31]. This is particularly true for populations that have strikingly different ancestry profiles, as is the case for Antioquia and Chocó. However, for this study, the general concordance seen between genetically inferred (predicted) phenotypic differences and the

observed differences for anthropometric traits, or known prevalence differences in the case of disease traits, supports the approach taken here (Table 6). It should be stressed that both trait-associated variant allele frequencies and PRS distributions overlap substantially between Antioquia and Chocó; in other words, predicted phenotypic differences vary along a continuum, with distinct group-specific averages in a minority of cases, as opposed to showing discrete values between populations. This is consistent with the expectation that the majority of genetic variation is found within rather than between human populations [115, 116].

Finally, it is important to note that detailed individual-level phenotypic information will be needed to more rigorously evaluate the implications of genetic divergence at trait-associated variants in diverse populations of the kind studied here. Fortunately, data of this kind are increasingly being generated by biobank collections around the world, via the combination of genetic profiles and detailed phenotypic information gleaned from participant surveys and electronic health records. Many of these biobanks – e.g., All of Us, BioMe, and the UK Biobank – include the kind of ancestrally diverse participant cohorts that can facilitate detailed investigations on the genetic basis of group-specific trait differences and health disparities.

The results reported here are distributed via a web-based platform that allows users to explore the extent of between-population divergence for individual trait-associated variants and PRS: http://map.chocogen.com

3.6 Supplemental information

3.6.1 PRS calculation and comparison among divergent populations

Our approach to PRS calculation and comparison between populations is characterized by three important choices: (1) the use of only significantly associated SNPs ($P<10^{-5}$) for PRS calculation, (2) the calculation of PRS that are unweighted by SNP effect sizes, and (3) the calculation of PRS without the use of linkage disequilibrium (LD) pruning or clumping. PRS were calculated in this way to facilitate comparisons of PRS distributions between divergent populations with distinct ancestry profiles. This conservative approach to PRS calculation and comparison is supported by (1) the lack of apparent systematic bias in PRS differences between populations (Figure 12), (2) the consistency of predicted risk differences between populations with previously reported trait and disease prevalence differences (Table 6), and (3) the consistency of PRS differences found when different ancestry SNP-association cohorts were used (Figure 13). Here, we provide additional justification for our approach to PRS calculation.

3.6.1.1 Top-SNP approach

There are many different ways to compute PRS, and the extremes are the "top-SNP" approach, where only highly significantly associated SNPs are used for PRS calculation and the "all-SNP" approach, whereas many SNPs as possible are used to calculate PRS. The top-SNP approach is more conservative as it relies only on robust associations. In contrast, the all-SNP approach derives additional resolution by capturing more of the genome-wide trait heritability. Most importantly, for our study, the all-SNP approach will also capture most or all of the population structure between divergent populations, whereas the top-SNP approach is not expected

to do so. In other words, when the all-SNP approach is used to compare *PRS* for divergent populations, such as the kind studied here, the resulting *PRS* are essentially guaranteed to show large differences between populations. This has been convincingly shown in a recent study, where increasing numbers of SNPs used for *PRS* yielded increasingly greater between-population differences, particularly for divergent populations [117].

Furthermore, recent work by Khera et al. suggests that the all-SNP approach only provides a marginal increase in prediction accuracy compared to the top SNP approach [118]. For example, a top-SNP PRS for coronary artery disease using 74 variants showed 79% accuracy compared to 81% accuracy when 6.6 million SNPs were used for PRS calculation. Similar marginal increases in accuracy between the top-SNP and all-SNP approaches were observed for the four other traits analyzed in the same study. These findings support both the utility of the top-SNP approach for cross-population PRS comparisons and its resolution for capturing the majority of variance in trait risk.

3.6.1.2 Unweighted *PRS*

PRS can be calculated as unweighted scores by simply summing the numbers of trait-associated effect alleles genome-wide, or they can be calculated as weighted scores, whereby each effect allele is weighted by its effect size (odds ratio or beta value). As with our previous studies [54, 59], we chose to use unweighted *PRS* here to facilitate the inclusion of SNP trait-associations from multiple studies. Effect sizes from different studies cannot be readily combined owing to differences in study cohorts, including cohort size, allele frequencies, and population structure. Furthermore, since effect sizes represent SNP heritability estimates, which are dependent on the particular cohort that is being studied, it does not make sense to attempt to normalize effect sizes across studies. Meta-analyses can perform this kind of normalization, as they have access to

individual-level phenotype data, but we do not have access to data of that kind data here. Finally, the use of multiple studies allowed us to ascertain as many trait-associated SNPs as possible, which is particularly important given our choice of the conservative top-SNP approach that is limited to significantly associated SNPs.

3.6.1.3 No linkage disequilibrium (LD) pruning

PRS are often calculated using linkage disequilibrium pruning or clumping, whereby only a single SNP from any given LD block is used for PRS calculation. We opted not to use LD pruning or clumping here because (1) we use a top-SNP approach for PRS calculation, and (2) the two populations under study have a highly divergent LD structure. The top-SNP approach means that we are using a relatively small number of SNPs per population, and the divergent LD structure means that different subsets of this small number of SNPs would likely be removed from each population if LD pruning were used. For example, LD pruning here may remove SNPs that are population-private (that is, a SNP that appears in Chocó but not Antioquia) or whose LD patterns are discordant between the two populations (i.e., the SNP is in moderate LD in Antioquia but low LD in Chocó). This would severely mitigate our ability to compare PRS between populations. An alternative approach, as discussed previously, would be to use a very large number of variants together with LD pruning for PRS computation, i.e., essentially covering most of all of the LD blocks in the genome for the PRS, as has been done in a number of studies. However, when tens- or hundreds-of-thousands, or even millions, of SNPs are used for PRS calculation between divergent populations, then the PRS is essentially modeling the population structure between the populations. In this case, all traits will be highly divergent if you are comparing divergent populations. Our approach to PRS calculation without LD pruning provides for both additional

resolution, in terms of the numbers of SNPs available for analysis, and more direct comparisons between divergent populations. Several studies, including our work, have shown that *PRS* calculated with and without LD pruning, do not show big differences [119, 120].

Table 1. Human populations analyzed in this study.

1KGP = 1000 Genomes Project *Phase III*

Dataset	Year	Population Name	Short	*n*
Colombian Populations				
Medina et al	2016	Chocoano in Quibdó, Colombia	CHG	94
1KGP	2015	Colombian in Medellin, Colombia	CLM	94
	Continental reference populations			
1KGP	2015	Yoruba in Ibadan, Nigeria	YRI	108
1KGP	2015	Iberian populations in Spain	IBS	107
1KGP	2015	Peruvian in Lima, Peru	PEL	85
Reich et al	2012	Embera in Colombia	Embera	5
Reich et al	2012	Quechua in Peru	Quechua	40
Reich et al	2012	Zapotec in Mexico	Zapotec	43

Table 2. Bioinformatics methods used in this study.

Software	**Use**	**Access**
PLINK version 1.9	SNP quality control, merging, and pruning	https://www.cog-genomics.org/plink2/
ShapeIT version 2.r837	SNP phasing	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
ADMIXTURE version 1.3.0	Continental genetic ancestry inference	http://software.genetics.ucla.edu/admixture/
IMPUTE2 version 2.3.2	Genome variant imputation	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Database	**Use**	**Access**
NHGRI-EBI GWAS Catalog	SNP trait associations and polygenic trait scores	https://www.ebi.ac.uk/gwas/ (accessed December 2018)

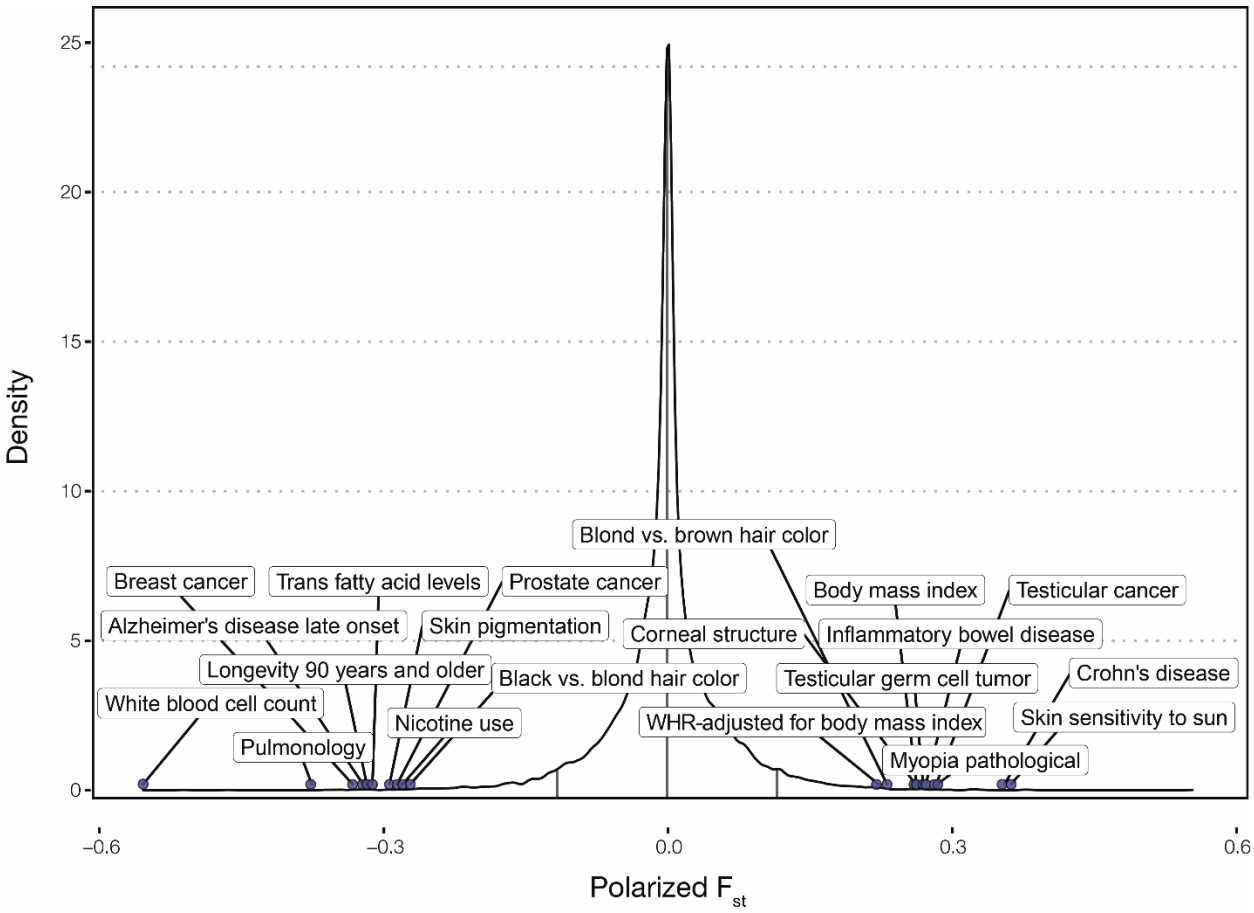

Figure 11. Distribution of polarized F_{ST} values between Antioquia and Chocó.

F_{ST} values were polarized to facilitate comparison between Antioquia (positive) and Chocó (negative). Highly differentiated alleles selected for Figure 2 are annotated (blue points). SNP F_{ST} and P-values are in Table 5.

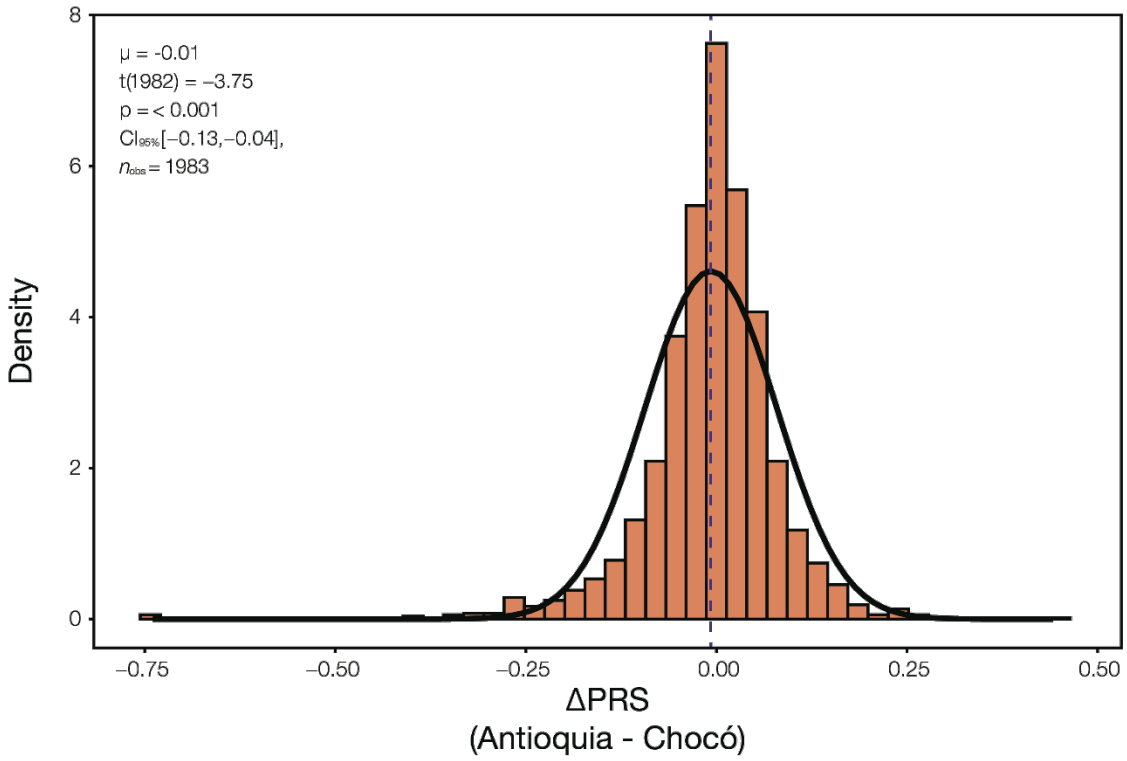

Figure 12. Distribution of PRS differences between Antioquia and Chocó

A histogram of the observed PRS differences is shown along with the corresponding smoothed density distribution. The mean difference value (μ) is indicated with a dashed line.

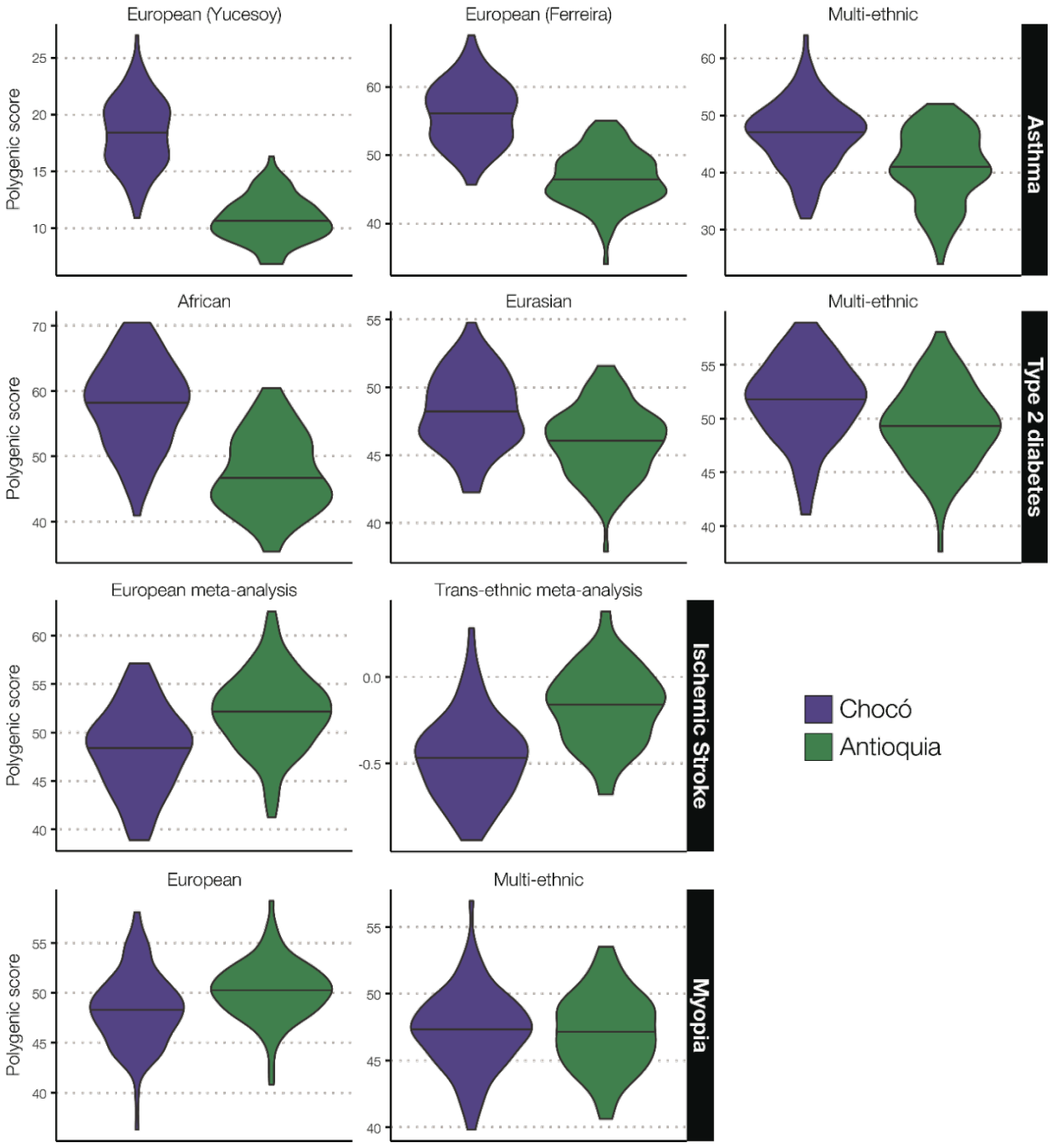

Figure 13. Effect of GWAS discovery population ancestry on PRS

Four selected traits from Figure 2 and Figure 3 were further analyzed with respect to GWAS discovery population ancestry, with two traits chosen for both Antioquia (ischemic stroke, myopia) and Chocó (asthma, type 2 diabetes). Trends in populations PRS distributions are similar regardless of GWAS ancestry.

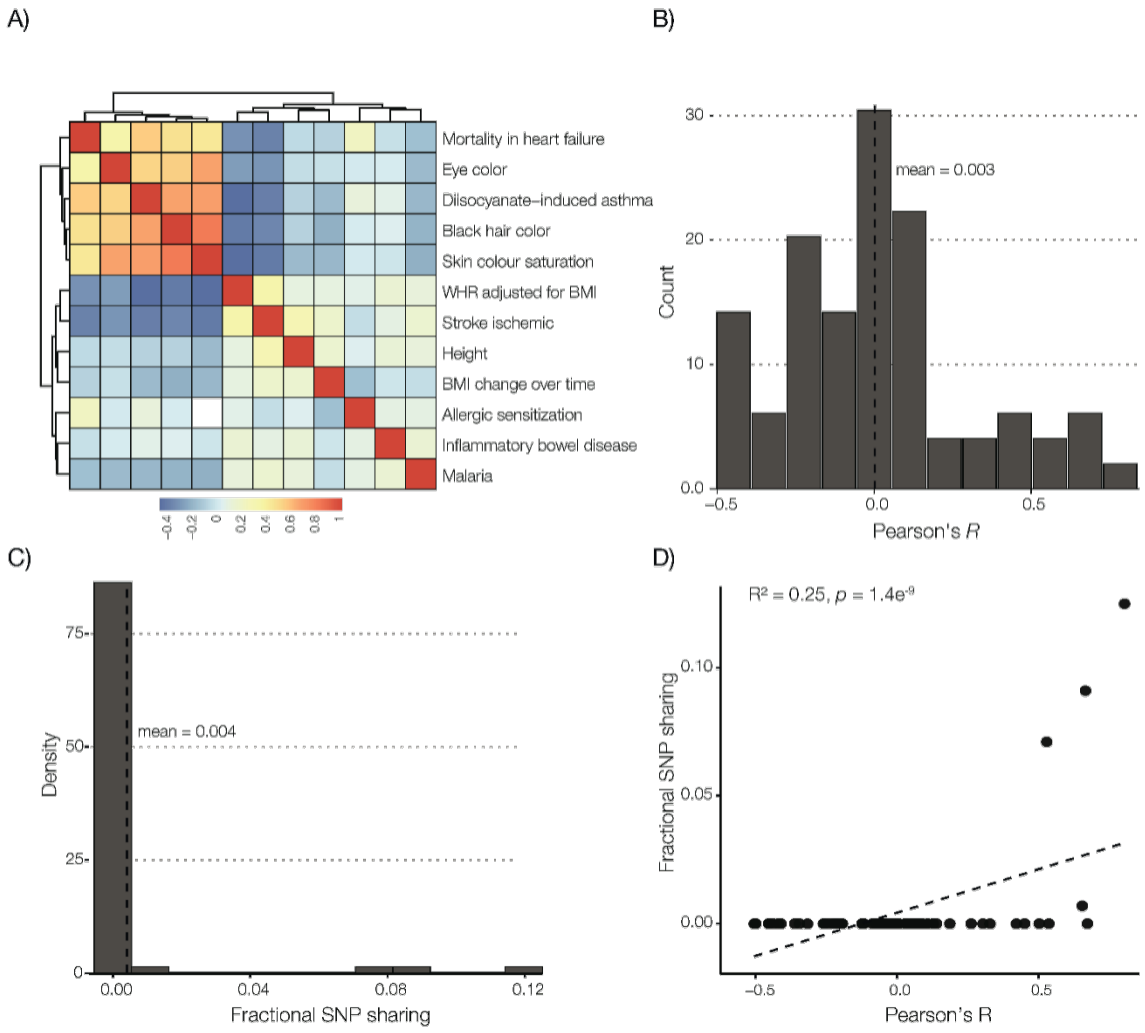

Figure 14. Correlations and SNP overlap among PRS

(A) All-by-all pairwise correlations between PRS from Figure 3B, hierarchically clustered by their Pearson correlations. (B) Distribution of Pearson correlation values, excluding self-correlation (n=132, mean=0.003). (C) Distribution of the fraction of SNPs shared between PRS (common SNPs in intersection/union of SNPs between scores). (D) Relationship between the fraction of SNPs shared between traits (y-axis) and the PRS correlation between traits (x-axis).

CHAPTER 4. INFLUENCE OF GENETIC ANCESTRY AND SOCIOECONOMIC STATUS ON DIABETES IN THE DIVERSE COLOMBIAN POPULATIONS OF CHOCÓ AND ANTIOQUIA

4.1 Abstract

Differences in genetic ancestry and socioeconomic status (SES) among Latin American populations have been linked to health disparities for a number of complex diseases, such as diabetes. We used a population genomic approach to investigate the role that genetic ancestry and socioeconomic status (SES) play in the epidemiology of type 2 diabetes (T2D) for two Colombian populations: Chocó (Afro-Latino) and Antioquia (Mestizo). Chocó has significantly higher predicted genetic risk for T2D compared to Antioquia, and the elevated predicted risk for T2D in Chocó is correlated with higher African ancestry. Despite its elevated predicted genetic risk, the population of Chocó has a three-times lower observed T2D prevalence than Antioquia, indicating that environmental factors better explain differences in T2D outcomes for Colombia. Chocó has substantially lower SES than Antioquia, suggesting that low SES in Chocó serves as a protective factor against T2D. The combination of lower prevalence of T2D and lower SES in Chocó may seem surprising given the protective nature of elevated SES in many populations in developed countries. However, low SES has also been documented to be a protective factor in rural populations in less developed countries, and this also appears to be the case when comparing Chocó to Antioquia.

4.2 Background

With ongoing economic development and the lifestyle changes that accompany increased standards of living, the primary disease burden in Latin America is shifting from infectious to non-communicable, complex diseases[121]. In fact, complex common diseases such as heart disease, cancer, and diabetes already account for the majority of the morbidity and mortality in the region[122]. Complex multifactorial diseases of this kind are associated with the effects of multiple genetic loci combined with a variety of environmental factors, such as diet, lifestyle, and exposure to toxins. The burden of complex disease is not evenly distributed within or between countries in Latin America; genetic and environmental differences among Latino populations often lead to pronounced health disparities[123]. Furthermore, population health disparities in Latin America tend to have a disproportionate impact on vulnerable Native American and Afro-Latino communities[124].

Diabetes mellitus is a complex multifactorial disease characterized by both a very high disease burden and strikingly disparate impacts among distinct populations in the Americas. For example, type 2 diabetes (T2D) has a substantially higher prevalence in both Native Americans and African-Americans compared to European-Americans in the United States (US)[125-130]. The higher prevalence of T2D in these populations has been associated with both genetic and environmental factors. Genetic risk for T2D is correlated with both increased Native American and African ancestry[131-134], and low socioeconomic status (SES) has also been widely associated with increased T2D prevalence in Native American and African-American populations[135-137].

Latin American populations are characterized by substantial genetic admixture – with predominant ancestry contributions from Europe, the Americas, and Africa – owing to historical

patterns of migration, conquest, and slavery[138]. Colombia has among the highest levels of three-way genetic admixture seen for any Latin American country[74, 76] and is home to a large population of Afro-descendants[73, 80, 139]. Estimates for the size of the Afro-Colombian population range from 9-20 million, making it the second-largest population of its kind in Latin America after Brazil. The collaborative ChocoGen research project was conceived to study the genetic heritage of the Afro-Colombian population from the administrative department (*i.e.*, state) of Chocó, located along Colombia's Pacific coast (http://www.chocogen.com)[80, 139]. The ChocoGen project has the joint aims of (1) characterizing the genetic ancestry of the population of Chocó, and (2) exploring the relationship between genetic ancestry and determinants of health and disease in the region.

The objective of the study was to evaluate the contributions of genetic ancestry and environmental factors to population health disparities in Chocó, and we addressed this issue here via a population genomic study of the genetic risk and the observed prevalence of T2D. Our efforts towards this end involve a comparison between the populations of Chocó and the neighboring state of Antioquia, which borders Chocó to the east (Figure 19). Despite their proximity, Chocó and Antioquia have very distinct demographic and economic profiles. According to the 2005 Colombian census, the population of Chocó was 82% Afro-Colombian, 13% Native American and 5% European/Mestizo, whereas Antioquia was 93% European/Mestizo, 7% Afro-Colombian and ~0.1% Native American[140]. The population of Chocó is considered to be particularly vulnerable, with high levels of poverty and low measures of economic development across several indices compared to Antioquia. We chose to focus our comparative study of genetic and health differences between Chocó and Antioquia on T2D for several reasons: (1) its high disease burden, (2) its known contribution to population health disparities, and (3) the relative wealth of knowledge

regarding its underlying genetic architecture. We set out to assess whether and how genetic and environmental differences between these two very distinct regions may manifest themselves with respect to population-specific levels of T2D genetic risk and/or differences in observed prevalence for the disease.

4.3 Materials and Methods

4.3.1 Genome sequence and genotype data sources

Whole genome sequence data and whole genome genotype data were analyzed in order to infer the genetic ancestry and admixture profiles for the Colombian populations of Chocó and Antioquia (Table 4). Whole genome genotypes for 94 individuals from Chocó were characterized as part of the ChocoGen research project (http://www.chocogen.com/) as previously described[80]. Sample donors from the ChocoGen project signed informed consent documents indicating their understanding of the potential risks of the project, along with how their data would be handled and how their identity would be protected. Collection, genotyping, and comparative analyses of human DNA samples were conducted with the approval of the ethics committee of the Universidad Tecnológica del Chocó[80].

All of the other data used for the analysis described here corresponds to publicly available and de-identified genome sequences or genotypes. Publicly available whole genome sequences for 94 individuals from Medellín, Antioquia were characterized as part of the 1000 Genomes Project (1KGP)[82]. Whole genome sequences from several additional admixed American populations were taken from the 1KGP for analysis: Utah residents with European ancestry (n=99), African ancestry individuals from the Southwest US (n=61), and a Peruvian population from Lima, Peru (n=85).

Genome sequence and genotype data were also sampled from putative ancestral populations corresponding to the three major continental regions that are known to contribute to genetic admixture in Colombia[73, 74, 76, 138]: Africa, Europe, and the Americas. African ancestry was inferred using whole genome sequences for a Yoruba population from Ibadan, Nigeria (n=108), and European ancestry was inferred using whole genome sequences for an Iberian population from Spain (n=107), both of which were characterized as part of the 1KGP. Whole genome genotypes for three Native American populations – Embera from Colombia (n=5), Quecha from Peru (n=40), and Zapotec from Mexico (n=43) – were taken from a dataset collected as part of a previous study on Native American genetic ancestry[83].

4.3.2 Genetic ancestry and admixture analysis

Whole genome sequence data and whole genome genotype data were merged using the program PLINK[141], and the resulting merged single nucleotide polymorphism (SNP) dataset was pruned in order to remove SNPs that are in linkage disequilibrium ($r \geq 0.05$). This resulted in a final dataset of 220,724 SNPs across 736 individual genome samples. Pairwise genomic distances between individuals were calculated as allele sharing distances between all pairs of merged/pruned SNP sets, also using PLINK. The pairwise allele sharing distance matrix was reduced to two-dimensions with principal component analysis (PCA) using the prcomp function from the R package for statistical computing[142] (Figure 15A). Ancestry fractions – African, European, and Native American – were calculated for each genome from Chocó and Antioquia using the program ADMIXTURE[87], with global reference populations (Table 4) and K=3 clusters corresponding to each of the major continental ancestry groups (Figure 20 and Figure 15B).

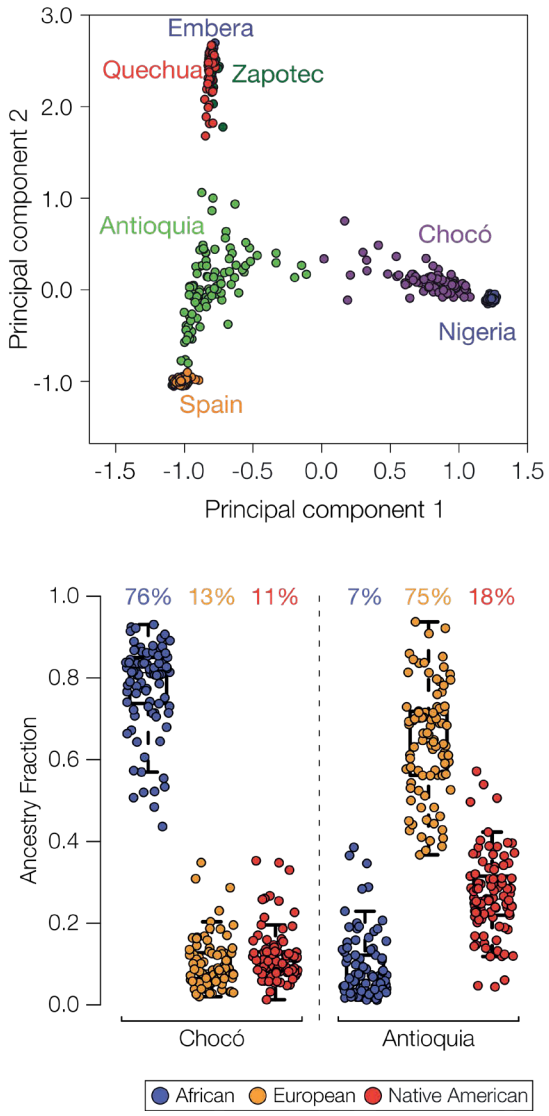

Figure 15. Genetic ancestry of the individuals from Chocó and Antioquia analyzed here.

(A) Principal components analysis (PCA) plot representing the pairwise distances among individual genomes from the admixed Colombian populations of Chocó and Antioquia along with putative ancestral source populations from Africa (Nigeria), Europe (Spain) and the Americas (Embera, Quechua, and Zapotec). (B) Box-plot distributions of the ancestry fractions for individuals from Chocó and Antioquia. The population-average values of African (blue), European (orange), and Native American (red) ancestry are shown above the distributions. Type 2 diabetes genetic risk calculation

The underlying genetic architecture of type 2 diabetes (T2D) was assayed from a series of 29 case-control genome-wide association studies (GWAS). T2D SNP association data from these studies were taken from the NHGRI-EBI GWAS catalog[143]. Individual SNP entries from the GWAS catalog were considered to be significantly associated with T2D if (1) the SNP association was uncovered via a case-control study based on at least 100,000 genotyped SNPs, (2) the SNP had the strongest association seen for its genomic locus, and (3) the SNP showed a genome-wide T2D association *P*-value<$1.0e^{-5}$. This yielded a set of 165 T2D-associated SNPs, and for each SNP, the identity of the risk allele (*i.e.*, specific nucleotide variant) linked to T2D was taken from the study where it was reported.

Imputation was performed on whole genome genotype data from the ChocoGen project in order to facilitate direct comparison of genome-wide T2D risk scores computed for datasets from Chocó (genotypes) and Antioquia (genome sequences). Before imputation, the whole genome genotypes of individuals from Chocó were phased using the program SHAPEIT[144, 145] using the 1KGP phase 3 haplotype reference panel. The phased whole genome genotypes from Chocó, consisting of 522,458 SNPs per individual, were then imputed using the program IMPUTE2[146-148] with the 1KGP phase 3 haplotype reference panel[145]. This process resulted in the imputation of 35,056,488 additional SNPs across all samples. The accuracy of the imputation process was evaluated by comparing the genetic ancestry relationships between individuals from Chocó, computed before and after imputation, and a panel of global reference populations. The observed genetic ancestry relationships for the individuals from Chocó are virtually identical before and after imputation, in support of the accuracy of the imputation process (Figure 21).

For each T2D-associated SNP, a log odds ratio (OR) was used to compute the relative genetic risk of T2D for Chocó compared to Antioquia (Figure 16**A**):

$$OR = Ln\frac{RA_{CHO}/NRA_{CHO}}{RA_{ANT}/NRA_{ANT}} \quad \textbf{(5)}$$

where RA_i and NRA_i are the risk allele frequency and non-risk allele frequency, respectively, in population i (CHO=Chocó and ANT=Antioquia). A meta-analysis was conducted to evaluate the joint effect of all 165 T2D-associated SNPs on the relative genetic risk of T2D in Chocó versus Antioquia using the metafor package in R[149]. 95% confidence intervals for the individual SNP and meta-analysis OR values were computed using both fixed- and random-effects models. The fixed- and random-effects models were both computed with moderators via linear (mixed-effects) models.

T2D polygenic risk scores (PRS) for individual genomes were computed as the unweighted, normalized sum of the number of risk alleles for all 165 T2D-associated SNPs (Figure 17**,** Equation (1). SNP association effect sizes were not used to weight the T2D PRS values owing to the fact that the T2D associated SNPs analyzed here were taken from different studies, and the effect size values among studies are not directly comparable.

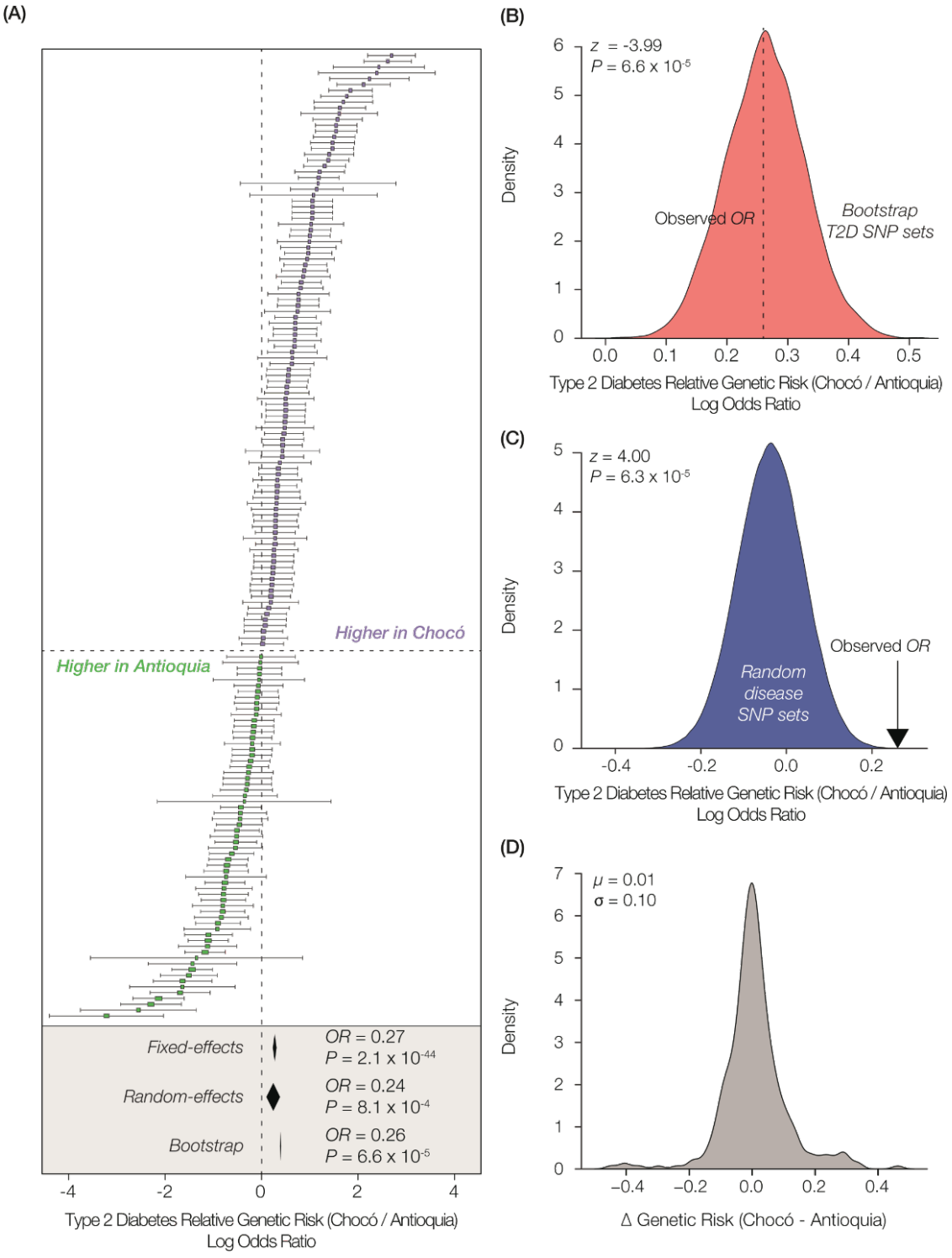
(A)
Higher in Chocó
Higher in Antioquia
Fixed-effects
OR = 0.27
P = 2.1 x 10-44
Random-effects
OR = 0.24
P = 8.1 x 10-4
Bootstrap
OR = 0.26
P = 6.6 x 10-5
-4
-2
0
2
4
Type 2 Diabetes Relative Genetic Risk (Chocó / Antioquia)
Log Odds Ratio
(B)
z = -3.99
P = 6.6 x 10-5
Density
Observed OR
Bootstrap
T2D SNP sets
0.0
0.1
0.2
0.3
0.4
0.5
Type 2 Diabetes Relative Genetic Risk (Chocó / Antioquia)
Log Odds Ratio
(C)
z = 4.00
P = 6.3 x 10-5
Density
Observed OR
Random
disease
SNP sets
-0.4
-0.2
0.0
0.2
Type 2 Diabetes Relative Genetic Risk (Chocó / Antioquia)
Log Odds Ratio
(D)
μ = 0.01
σ = 0.10
Density
−0.4
−0.2
0.0
0.2
0.4
Δ Genetic Risk (Chocó - Antioquia)

Figure 16. Relative genetic risk for type 2 diabetes (T2D) and genetic ancestry in Chocó versus Antioquia.

(A) The relative genetic risk of T2D in the two Colombian populations is shown as log odds ratios (OR) – Chocó/Antioquia – of the risk versus non-risk allele frequencies for 165 T2D-associated SNPs. The formula for calculating OR values is shown in the Methods subsection 'Type 2 diabetes genetic risk calculation' (formula 1). OR values > 0 indicate greater risk in Chocó (purple), whereas OR values < 0 show greater risk in Antioquia (green). 95% confidence intervals (CI) for individual SNP OR values are shown. The diamonds below the plot show OR values (±95% CI) corresponding to fixed- and random-effects meta-analysis of all 165 T2D-associated SNPs as well as the mean OR value from the bootstrap analysis; P-values indicating the statistical significance level of the three meta OR values are shown. (B) The observed OR value for the relative genetic risk of T2D (Chocó/Antioquia) is compared to a bootstrap distribution of OR values based on random sampling with replacement from the set of T2D-associated SNPs. The values of z and P for a z-test comparing the distribution of bootstrap T2D SNP OR values to 0 are shown. (C) The observed OR value for the relative genetic risk of T2D (Chocó/Antioquia) is compared to a null distribution of expected OR values for randomly simulated SNP sets of the same size as the T2D-associated SNP set. The values of z and P for a z-test comparing the observed and expected T2D SNP OR values are shown. (D) The distribution of genetic risk score (PRS) differences (Chocó-Antioquia) for 324 diseases is shown along with the mean and standard deviation values for the distribution.

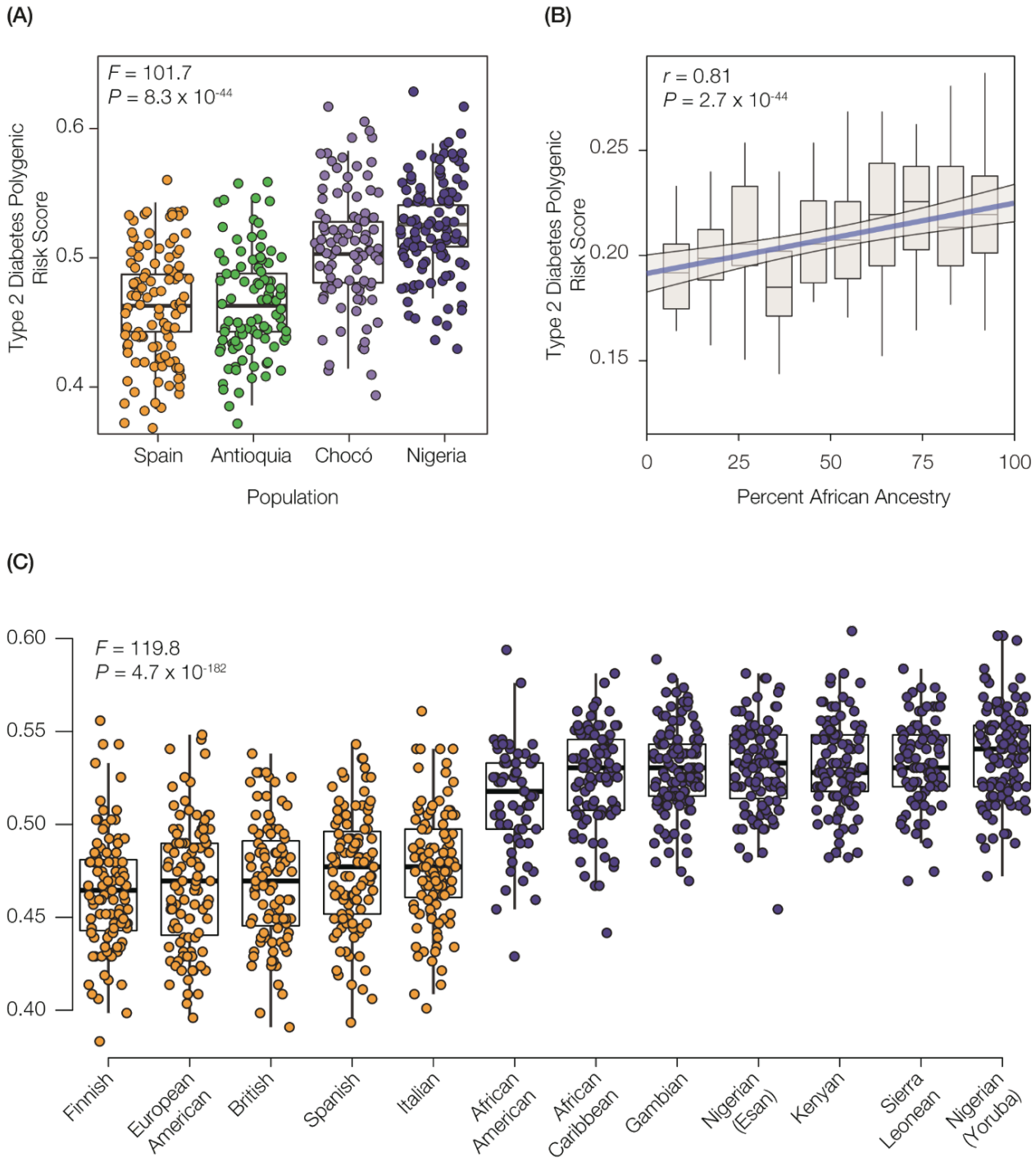

Figure 17. Genetic ancestry and predicted risk for T2D.

(A) Box-plot distributions of individuals' T2D polygenic risk scores are shown for four populations: Spain (orange), Antioquia (green), Chocó (purple), and Nigeria (blue). The values of F and P for an ANOVA test comparing the mean values of the distributions are shown. (B) Regression of T2D polygenic risk scores (y-axis) against the percent African ancestry for genome sequences from Colombia and the US (x-axis). Box plots are shown for decile bins, and the linear

trend line is shown in blue with 95% CI in gray. The values of r and P for the Pearson correlation coefficient of the regression are shown. (C) Box-plot distributions of individuals' T2D polygenic risk scores are shown for five European populations (orange) and seven African populations (blue). The values of F and P for an ANOVA test comparing the mean values of the distributions are shown.

4.3.3 Genetic risk calculation controls

A series of controls was performed to check for systematic biases in the frequencies allelic variants used to compare genetic risk scores between populations. (1) Bootstrap: random sampling with replacement from the 165 T2D-associated SNPs was used to create 10,000 replicate SNP sets, each of which was used for genetic risk OR calculation and meta-analysis as described above. The resulting distribution of bootstrap meta-analysis OR values was compared to the observed value for the T2D SNP set to evaluate how outliers may affect T2D genetic risk calculation and comparison between populations (Figure 16B). (2) Random disease-associated SNP sets: random sampling of T2D size-matched (n=165) disease-associated SNP sets from the NHGRI-EBI GWAS catalog was used to create 500,000 replicate SNP sets, each of which was used for genetic risk OR calculation and meta-analysis as described above. The resulting distribution of random disease-associated SNP set meta-analysis OR values was compared to the observed value to evaluate whether systematic biases in disease-associated allele frequencies between populations may affect the comparison of genetic risk (Figure 16C). (3) Disease genetic risk comparisons: SNP disease-associations from the NHGRI-EBI GWAS catalog were mined to compare polygenic risk scores (*PRS*), as described above for T2D, for 324 diseases between Chocó and Antioquia in order to assess whether there is any systematic bias in disease genetic risk score computation between the two populations (Figure 16D).

4.3.4 Diabetes prevalence and socioeconomic status (SES) data sources

Data on age-adjusted diabetes prevalence per 100,000 inhabitants for the Colombian administrative departments (*i.e.*, states) was taken from three database sources: (1) Cuenta de Alto Costo (https://cuentadealtocosto.org/), (2) Observatorio de Diabetes de Colombia (http://www.odc.org.co/), and (3) the Sistema Integral de Información de la Protección Social databases (https://www.minsalud.gov.co/salud/Paginas/SistemaIntegraldeInformaci%C3%B3nSISPRO.aspx) (Figure 18). Data on SES indicators was collected from the Departamento Administrativo Nacional de Estadística (DANE)[94] and Instituto Colombiano de Bienestar Familiar[95] (Table 2).

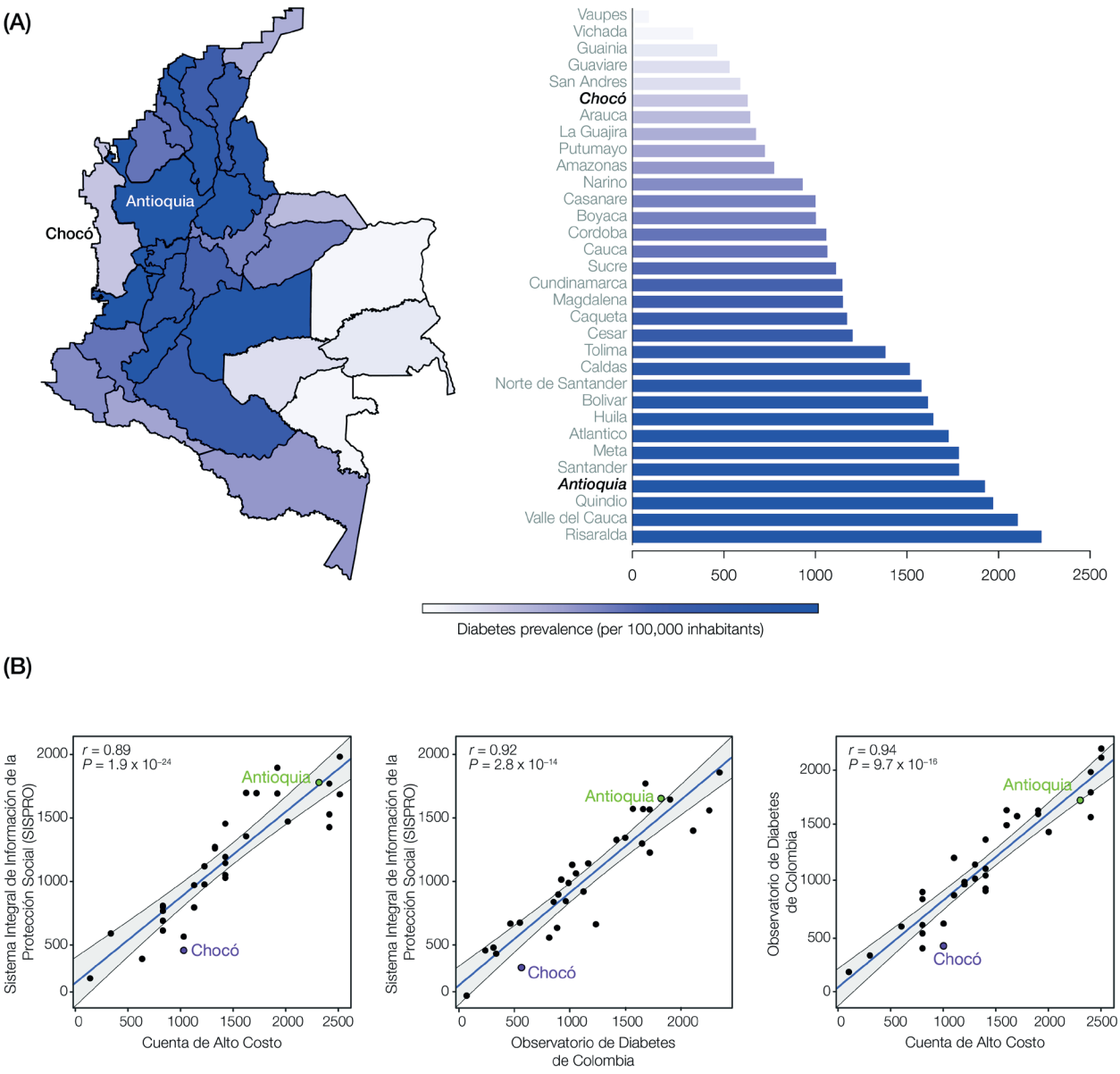

Figure 18. Prevalence of diabetes in Colombia.

(A) Age-adjusted diabetes mellitus prevalence per 100,000 inhabitants. (B) Comparison of Colombian diabetes state-by-state prevalence estimates taken from the three different database sources. Regression plots for all three possible pairwise comparisons between the different databases are shown, with the values for Chocó and Antioquia indicated. For each regression, the Pearson correlation r-value is shown along with the P-value significance level.

4.4 Results

4.4.1 Comparative genetic ancestry

Here and elsewhere[74, 80], we characterized the genetic heritage of Chocó and Antioquia with respect to their populations' ancestry proportions derived from Africa, Europe, and the Americas. To do so, whole genome genotypes characterized for donors from Chocó, along with publicly available whole genome sequences from Antioquia, were compared to genomes from putative ancestral source populations collected from a variety of sources (Table 4). Details of the approaches we used for all comparative genomic analyses can be found in the Methods section. Pairwise genomic distances projected onto two dimensions group individuals from Chocó with an African population from Nigeria, whereas individuals from Antioquia group most closely with a European population from Spain (Figure 15A). Nevertheless, both populations show visual evidence of substantial admixture among the three major continental population groups on this same plot. The inferred continental genetic ancestry fractions for Chocó and Antioquia are also largely consistent with the states' demographic profiles, which were gleaned from self-reported ethnicity, with Chocó having predominantly African ancestry and Antioquia having mainly European ancestry. Admixture analysis revealed that the population of Chocó has 76% African, 13% European, and 11% Native American ancestry, whereas Antioquia has 75% European, 18% Native American and 7% African ancestry (Figure 15B and Figure 20).

4.4.2 Comparative T2D genetic risk

We asked whether the differences in genetic ancestry between Chocó and Antioquia are related to population-specific genetic risk for diabetes by comparing the distributions of known T2D risk alleles for the two populations using the previously described genomic datasets. T2D

risk alleles for a total of 165 single nucleotide polymorphisms (SNPs) were mined from a collection of 29 T2D genome-wide association studies (GWAS) (Table 3). Population-specific frequencies of the risk and non-risk alleles for each T2D-associated SNP were measured and used to calculate a log odds ratio (*OR*) that expresses the relative genetic risk of T2D for the two populations: Chocó/Antioquia. Log odds ratios were used to provide a statistical framework to measure the T2D risk contributions of individual SNPs and to allow for a meta-analysis that considers the additive genetic risk contribution of all SNPs together. Details of this approach are provided in the Methods section. The majority of T2D associated SNPs show higher risk allele frequencies in Chocó compared to Antioquia, pointing to a relatively higher genetic risk of T2D in the population of Chocó (Figure 16**A**). Ninety-one (91) individual SNPs show significant differences in risk versus non-risk allele frequencies in Chocó compared to Antioquia; 62 (68%) of those SNPs reflect significantly greater T2D genetic risk in Chocó compared to only 29 (32%) with a higher risk in Antioquia. When all of the T2D-associated SNPs are considered together using meta-analysis, Chocó shows a significantly greater population-wide genetic risk for T2D than Antioquia. Chocó/Antioquia T2D meta-analysis *OR* values, along with their 95% confidence intervals, were computed using both fixed and random effect models as well as via bootstrap analysis. All three approaches show significantly higher T2D genetic risk in Chocó compared to Antioquia (Figure 16A).

We performed a series of controls in an effort to ensure that the difference observed for T2D genetic risk between Chocó and Antioquia cannot be attributed to any systematic bias in the SNP allele frequencies of the two populations (Methods). First, we used a bootstrap analysis of the T2D SNP set to evaluate the signal-to-noise ratio in the data. In particular, we wanted to assess

whether the observed difference in T2D genetic risk between Chocó and Antioquia may be due to a few outlier SNPs (Figure 22). Sampling with replacement from the set of T2D-associated SNPs was used to generate 10,000 replicate T2D SNP sets, each of which was used to calculate a meta-analysis *OR* value. The distribution of bootstrap replicate *OR* values is centered around observed *OR* value, and the mean bootstrap *OR* value is significantly greater than 0 (Figure 16B; z=-3.99, P=6.6 x 10^{-5}). The results of the bootstrap analysis are consistent with greater T2D genetic risk in Chocó. They indicate that the signal in the data, based on the individual SNP *OR* values, is robust to sampling noise.

We next addressed whether the observed difference in T2D genetic risk can be attributed to a systematic bias in the allele frequencies for disease-associated SNPs between the two populations. This is particularly relevant given the fact that the vast majority of GWAS are conducted on populations of European ancestry, more similar to what is seen for Antioquia. In fact, it has recently been shown that attempts to compare genetic risk between populations with divergent ancestry profiles can be confounded by demographic factors that yield differences in the overall frequencies of risk alleles; effects of this kind can, in turn, lead to systematic biases in population-specific genetic risk estimates[31]. We attempted to control against this possibility using the two approaches described below.

We developed a simulation-based approach in order to control for the possible effects of demographic history on estimates of population-specific T2D genetic risk for Chocó and Antioquia. If the apparent elevated genetic risk for T2D in Chocó reflects a bias in the relative frequencies of disease-associated SNPs, perhaps owing to increased African ancestry of the population, then we would expect to see an overall shift to higher estimated disease risk for Chocó

compared to Antioquia. To evaluate this possibility, 500,000 SNP sets of the same size as the set of T2D-associated SNPs were randomly simulated from a collection of disease-associated SNPs taken from the NHGRI-EBI GWAS catalog[143]. For each of these random SNP sets, a meta-analysis of the SNP relative genetic risk log odds ratios (Chocó/Antioquia) was performed, yielding a random meta-analysis log odds ratio value (*OR*). The null distribution of the resulting random meta-analysis *OR* values was then compared to the observed T2D relative genetic risk *OR* value for Chocó/Antioquia. Contrary to the expectations of the demographic bias model, Antioquia shows a higher overall relative genetic risk when ensembles of randomly sampled disease-associated SNP sets are analyzed (Figure 16**C**). In addition, the observed T2D relative genetic risk *OR* value for Chocó/Antioquia is significantly greater than the expected *OR* value based on the null distribution, further validating the observed elevated genetic risk for T2D in Chocó (z=4.0, P=6.3 x 10^{-5}).

In addition to the simulation-based approach described above, we also used disease-associated SNPs from the NHGRI-EBI GWAS catalog to compute the relative genetic risk between Chocó and Antioquia for 324 additional diseases. In this case, a systematic bias in the population-specific allele frequencies of disease-associated SNPs would be expected to reveal an overall elevation of disease genetic risk in one of the two populations. However, the distribution of the differences in predicted genetic risk for these diseases is centered very close to 0 and more or less symmetrical (Figure 16D); the mean genetic risk difference (Chocó - Antioquia) for these diseases is not significantly different from 0 (z=-0.1, P=0.92). Taken together, these three controls suggest that the observed difference in T2D genetic risk for Chocó versus Antioquia cannot be attributed to any systematic bias in disease-associated allele frequencies between the two populations.

4.4.3 Genetic ancestry and T2D risk

Considering their respective ancestry profiles, the higher T2D genetic risk that we observe for the population of Chocó compared to Antioquia is consistent with previous results showing a correlation between African genetic ancestry and T2D prevalence in the US[132]. We asked whether the elevated genetic risk of T2D in Chocó may also be related to greater African ancestry, and conversely, lower European ancestry, in Chocó compared to Antioquia. To do this, we computed polygenic T2D risk scores for individuals from Chocó and Antioquia along with individuals from their most closely related putative ancestral populations in Europe (Spain) and Africa (Nigeria). We applied a widely used approach that computes polygenic risk scores for individual genomes, or whole genome genotypes, based on the sum of risk alleles present across all associated SNPs[37, 150-152] (Methods). The Antioquia population has the lowest T2D genetic risk measured this way, followed by the Spanish population; however, the T2D genetic risk score distributions between these two populations are not significantly different (t=0.3, P=0.8; Figure 17A). Chocó has significantly greater T2D genetic risk than Antioquia (t=5.7, P=4.1 x 10^{-8}), and the Nigerian population has the highest overall risk (Figure 17A). Thus, the T2D genetic risk score distributions for these populations follow the increasing proportions of African ancestry and decreasing European ancestry, seen among them. We also performed a similar analysis of T2D genetic risk analyses for a pair of African-American and European-American populations from the US, and find the same patterns of elevated T2D genetic risk associated with African ancestry that we see for Colombia, consistent with previous results[132] (Figure 23). Finally, we show that the African ancestry percentages for individuals from Colombia and the US are positively correlated with their polygenic risk scores for T2D (r=0.81, P=2.7 x 10^{-44}; Figure 18B).

We further evaluated the relationship between genetic ancestry and T2D genetic risk worldwide by comparing five European populations to seven African populations (Figure 17C). All of the African populations have higher T2D genetic risk than the European populations, and the difference between the African versus European ancestry group T2D genetic risk averages is highly significant (t=33.9, P=1.4 x 10^{-164}). These results lend additional support to the association of African genetic ancestry with elevated T2D genetic risk.

4.4.4 Observed T2D prevalence

Given the elevated genetic risk for T2D in the Afro-Colombian population of Chocó, along with its association with African ancestry, we expected to see a substantially higher prevalence of diabetes in Chocó compared to Antioquia. Indeed, numerous studies report that African-Americans in the US have a far higher prevalence of T2D than European-Americans[128-130]. However, we were surprised to find that the reported prevalence of diabetes is, in fact, more than three times higher in Antioquia than in Chocó (Figure 18A). Averaging data from three separate epidemiological database sources, maintained by governmental and non-governmental organizations, shows Antioquia with an age-adjusted diabetes prevalence of 1.9%, which is the 4th highest out of 32 states in the country, compared to 0.6% for Chocó, which is ranked 27th. The large difference in diabetes prevalence observed for Chocó versus Antioquia is highly consistent across the three different Colombian epidemiological databases that we sourced (Figure 18B).

The far lower prevalence of diabetes in Chocó versus Antioquia, compared to what may be expected based on the genetic profiles of their populations, strongly suggests that environmental factors predominantly shape diabetes outcomes in the region. This would be consistent with several large cohort studies showing that environmental factors contribute substantially more to T2D than genetic factors[153-155], and the populations of Chocó and Antioquia do indeed occupy

very distinct environments. In particular, as previously stated, the population of Chocó has far lower overall SES compared to Antioquia (Table 2). For example, the per capita gross domestic product in Chocó is almost three times lower than that of Antioquia. Chocó also has lower levels of literacy, life expectancy, employment, and modern housing, along with higher dietary deficits of protein and calcium than Antioquia. Considered together, these factors give Chocó a human development index (HDI) of 0.73, ranked 31st out of 32 Colombian states, compared to an HDI of 0.85 for Antioquia, which ranks 4th in the country. Thus, it appears that even though low SES has been associated with the risk for T2D in numerous studies[156], in Chocó low SES somehow serves as a protective factor against T2D. This unexpected finding suggests that poverty may play a very different role in the etiology of complex disease, particularly for diabetes and perhaps other metabolic syndrome disorders, in Colombia compared to more developed countries in the Global North.

4.5 Discussion

Our study of the contributions of genetic ancestry and environmental factors to T2D prevalence in two divergent Colombian populations suggests that poverty can serve as a T2D protective factor in Colombia. The possibility that poverty in Chocó is an environmental protective factor against T2D, as opposed to a strong risk factor as seen for African-Americans in the US, may be attributed to the differing nature of poverty in developed countries compared to some parts of the developing world. Poverty in the US is associated with poor diet and other lifestyle factors that elevate T2D prevalence[136, 137, 157, 158]. However, poverty in Chocó, which is generally more extreme than what is found in the US, is actually associated with a diet that is protective against T2D, particularly when compared to Antioquia. The dietary staples of Chocó are fish, plantains, yucca, and rice; fish are readily available from the Atrato River and its tributaries, and

plantains and yucca are cultivated along the banks of this vast river system[159, 160]. Thus, the typical diet of Chocó is high in polyunsaturated lipids, such as omega-3 and omega-6 fatty acids, and fiber, both of which are known to mitigate T2D risk. In Antioquia, the main sources of protein are beef and pork, which are rich in both cholesterol and triglycerides formed by saturated fatty acids, known risk factors for T2D. In addition to the ready availability of fish in the region, SES in Chocó also impacts dietary choices in a way that is protective against T2D. In Quibdó, the capital of Chocó, one kilogram of meat costs 9000 Colombian pesos or approximately $3 US dollars; 10kg of fish from the Atrato River can be bought for the same amount, providing a week's worth of protein.

We also found Chocó and Antioquia to be distinct with respect to the prevalence of alcohol consumption and tobacco use, both of which have been implicated as environmental factors that influence T2D outcomes. A 2013 government survey on the consumption of psychoactive substances in Colombia found that Chocó had the highest prevalence of alcohol consumption for the country, with 44.6% of respondents reporting alcohol consumption over the past 30 days compared to 36.6% for Antioquia[161]. Since moderate alcohol consumption has been linked to reduced risk for the onset of T2D[162-164], this could represent an additional protective factor associated with the lifestyle in Chocó. Conversely, Antioquia was found to have higher tobacco use in the same survey, with 14.1% use over the last month compared to 6.6% for Chocó. Smoking is a known risk factor for T2D[165-167], pointing to yet another possible advantage of the lifestyle in Chocó with respect to T2D prevalence.

Another way to consider the discordant results that we observed for population-specific genetic risk versus the observed prevalence of diabetes in Colombia is through the lens of economic development as opposed to poverty *per se*. While the notion that poverty in Chocó

serves as a protective factor against diabetes was certainly unexpected to us, if we consider Chocó to be under-developed relative to Antioquia, then the environmental protective effect may not be as surprising. Indeed, as previously stated, the HDI for Chocó points to substantial under-development compared to the rest of the country, and the pyramid-shaped age distribution of Chocó is more consistent with what is seen in less developed countries; the narrower age distribution of Antioquia, on the other hand, resembles those of more developed countries (Figure 24). T2D has been considered to be a disease of the developed world, as it is generally more prevalent in industrialized than less-developed countries[168]. In fact, studies have shown precipitous increases in T2D prevalence in populations that have undergone rapid transitions to more developed economies[169]. The comparison of Chocó versus Antioquia may underscore the public health relevance of stark differences in economic development within a single country, albeit in a way that counterintuitively favors the less developed region.

It is also worth noting that Chocó is more rural and less urbanized than Antioquia. Chocó is relatively underpopulated, with a population density of 11 individuals per km^2 compared to 99 per km^2 for Antioquia. The rural setting of Chocó, along with the overall challenging conditions of its environment, is associated with a more physically active lifestyle compared to more modernized parts of the country[159, 160], highlighting yet another potentially protective factor against T2D. Interestingly, a recent study in India showed that low SES is simultaneously a risk factor for T2D in cities and a protective factor for T2D in more rural areas [170]. Thus, it may be the case that urban poverty in developing countries is more reminiscent of overall poverty in the developed world, in terms of risk for T2D, whereas the features of rural poverty in the developing world are distinctive and protective for T2D.

We explored the relationship between economic development and T2D for the entire country by comparing HDI levels to T2D prevalence estimates for all states. We observe a strong positive correlation between HDI and T2D prevalence across Colombia, with more developed regions of the country showing higher T2D prevalence estimates (Figure 25A). This finding is consistent with the notion that lower levels of development within the country can serve as a protective factor against T2D. However, it could also be taken to suggest the possibility that the lower prevalence of T2D in Chocó reflects a bias in disease reporting, owing to lower SES and accordingly reduced access to healthcare services. We evaluated this possibility by comparing prevalence estimates for 43 diseases between Chocó and Antioquia. A reporting bias for Chocó, based on reduced access to healthcare, would be expected to reveal itself as an overall reduction in prevalence estimates for numerous diseases. In fact, Chocó shows a greater prevalence for 24 diseases, compared to 19 for Antioquia, and the difference between the two is not statistically significant (Figure 25B). These results indicate that a reporting bias based on differential access to healthcare does not likely explain the lower prevalence of T2D observed for Chocó.

Genomic approaches to health care, while still in their infancy in the region, hold great promise for Latin America, especially as the public health burden continues to shift toward common complex diseases with at least partial genetic etiology. The distinction that we observe between population-specific genetic risk and observed prevalence of T2D provides lessons for the implementation of genomic approaches to personalized and precision medicine in Latin American countries such as Colombia. Caution should be taken when extrapolating results from studies in the Global North, where the vast majority of this kind of research is still conducted[16, 32, 33], to developing countries in Latin America. For instance, a commonly accepted environmental risk factor for many common diseases, such as SES, may have very different implications in Latin

America compared to the US. In addition, the public health value of dietary and lifestyle choices, which may have been historically dictated by poverty, should be recognized and incorporated into public health campaigns as countries in Latin America continue to experience rapid economic development and urbanization. A corollary of this suggestion would be to strategically avoid pitfalls of urbanization in the developed world, such as the increasingly sedentary lifestyle, the reliance on processed and fast foods as well as the emergence of so-called 'food deserts' in poor neighborhoods where it is exceedingly difficult, if not impossible, to access fresh and whole foods.

Finally, caution also needs to be exercised when extrapolating the results of studies on the genetic architecture of complex diseases between populations with distinct ancestry profiles[31]. Genetic associations discovered in one population may not replicate in a different population, and ancestry and admixture can have additional confounding effects on the expression of genetic variants. Nevertheless, as we have endeavored to show here, exploration of disease-associated variants in understudied populations can provide valuable insight into the joint contributions of genetics and environment to common complex diseases, which are an increasing public health threat to the developing economies of the Global South.

4.6 Supplementary Information

Figure 19. Relief map of Colombia showing the locations of the administrative departments (*i.e.*, states) of Chocó and Antioquia.

Map adapted from https://commons.wikimedia.org/wiki/File:Mapa_de_Colombia_(relieve).svg, edited to highlight the states of interest. The image file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license https://creativecommons.org/licenses/by-sa/3.0/deed.en.

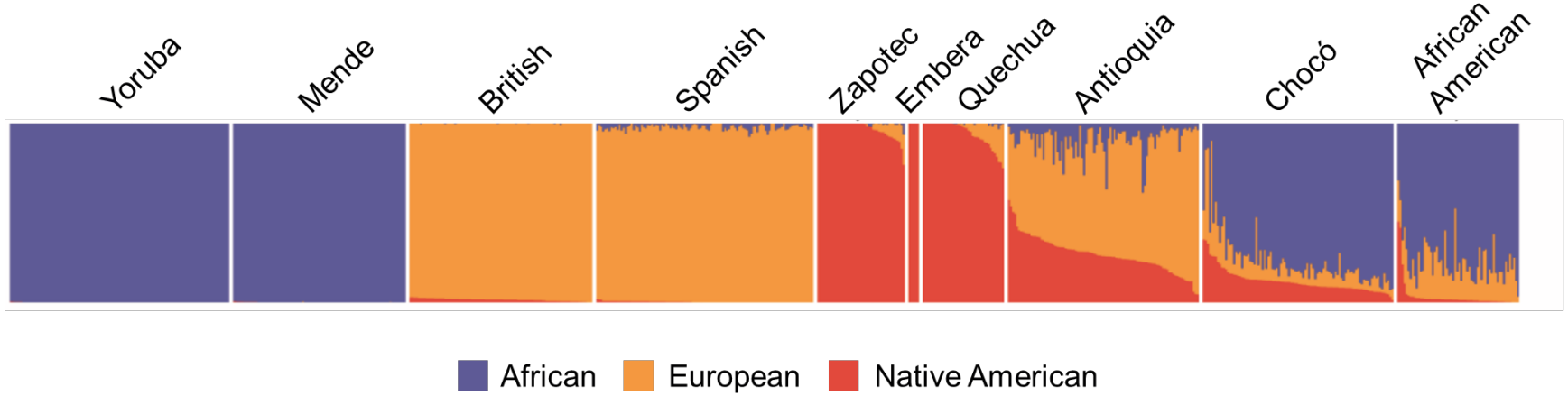

Figure 20. Admixture bar chart showing the percentage of African (blue), European (orange), and Native American (red) ancestry for the individuals from Antioquia Chocó analyzed here.

ADMIXTURE was run with K=3 clusters, corresponding to the three continental ancestry groups, using the global putative ancestral source populations shown here[80].

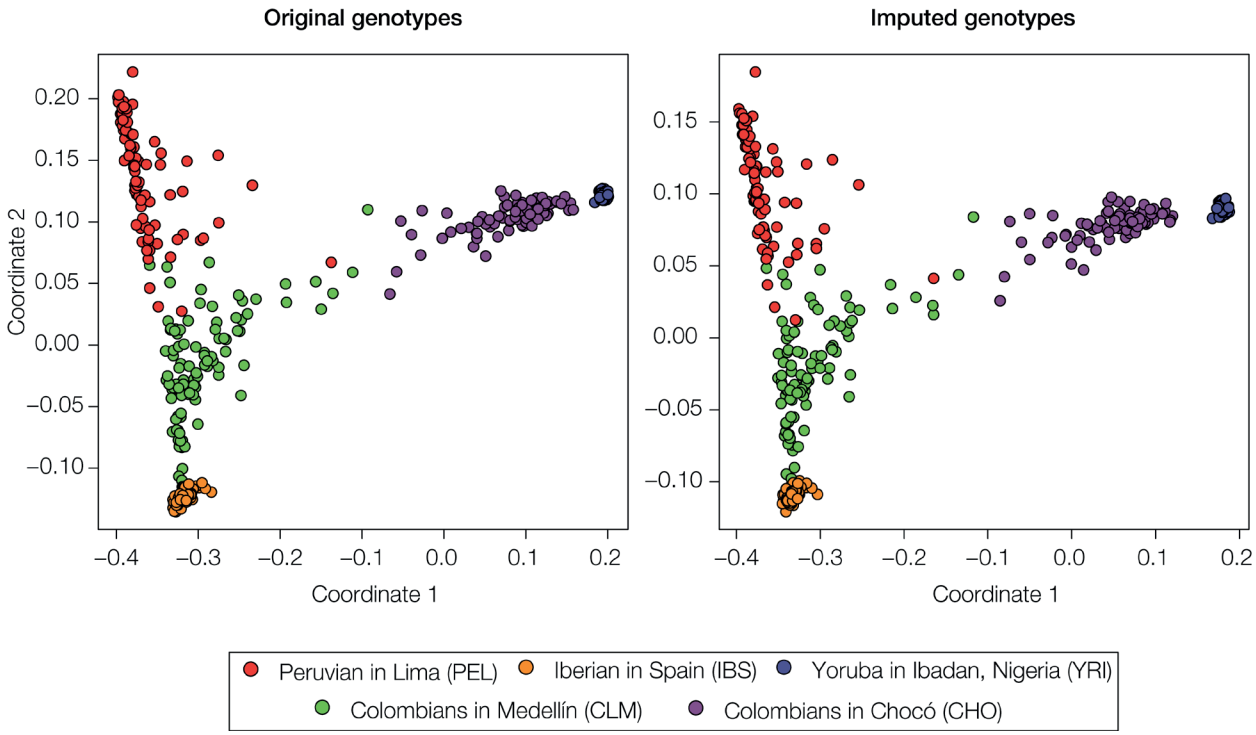

Figure 21. Validation of the SNP imputation process for the Chocó genotypes via comparison of genetic ancestry patterns before (original genotypes) and after (imputed genotypes) imputation.

Pairwise genomic distances between individuals from Chocó and a panel of global reference populations (see color key) characterized via whole genome sequencing as part of the 1KGP are shown before and after imputation.

Table 3. Type 2 diabetes (T2D) associated SNPs analyzed in this study.

165 T2D-associated SNPs, corresponding to 29 studies, were taken from the NHRGI-EBI GWAS database. [1]T2D-associated SNP identifier numbers, from the NCBI Single Nucleotide Polymorphism database (dbSNP https://www.ncbi.nlm.nih.gov/projects/SNP/). [2]Chromosomal locations for the T2D-associated SNPs, corresponding to the GRh37/h19 version of the human genome reference sequence. [3]Identity of the risk alleles (nucleotide variants) that are associated with a higher risk of T2D for each SNP. Risk allele identities are shown here for the positive DNA strand. [4]HGNC gene symbols for the genes considered to be influenced by the T2D-associated SNPs, according to each study. [5]NCBI PubMed identifier numbers for the publications where each T2D-associated SNP was reported.

rsID[1]	Chr[2]	Pos[2]	Risk Allele[3]	Gene(s)[4]	PubMed ID[5]
rs5945326	X	153634467	A	DUSP9	20581827
rs17106184	1	50444313	G	FAF1	24509480
rs11165354	1	91728765	A	TGFBR3	23209189
rs7542900	1	94604485	C	SLC44A3, F3	22238593
rs10923931	1	119975336	T	NOTCH2, ADAM30	18372903
rs2075423	1	213981376	G	PROX1	24509480
rs2820446	1	219575476	C	LYPLAL1	24509480
rs6426514	1	228744368	A	RHOU	23300278
rs12027542	1	233204408	A	PCNXL2	21490949
rs679992	1	241024982	T	intergenic	25102180
rs10190052	2	646674	C	TMEM18	24509480
rs11677370	2	3793830	T	intergenic	21490949
rs12613372	2	30845153	G	GALNT14, CAPN13	25102180
rs7578597	2	43505684	T	THADA	18372903
rs243088	2	60341610	T	BCL11A	24509480
rs243021	2	60357684	A	BCL11A	20581827
rs73954691	2	127663671	G	LIMS2	25483131
rs6723108	2	134722410	T	TMEM163	23209189
rs7560163	2	150781422	C	RBM43, RND3	22238593
rs3923113	2	164645339	A	GRB14	21874001
rs2943640	2	226228869	C	IRS1	24509480
rs17036101	3	12236345	G	SYN2, PPARG	18372903
rs13081389	3	12248301	A	PPARG	20581827
rs1801282	3	12351626	C	PPARG	17463246
rs6780569	3	23156993	G	UBE2E2	23945395
rs7612463	3	23294959	C	UBE2E2	24509480
rs831571	3	64062621	c	PSMD6	22158537
rs4607103	3	64726228	C	ADAMTS9	18372903
rs2063640	3	102484201	A	ZPLD1	21490949
rs11708067	3	123346931	A	ADCY5	22693455

Table 3 continued

rsID[1]	Chr[2]	Pos[2]	Risk Allele[3]	Gene(s)[4]	PubMed ID[5]
rs11717195	3	123363551	T	ADCY5	24509480
rs3773506	3	142712158	C	PLS1	21490949
rs7630877	3	179943530	A	PEX5L	21490949
rs4402960	3	185793899	T	IGF2BP2	17463246
rs1470579	3	185811292	C	IGF2BP2	20581827
rs6769511	3	185812502	C	IGF2BP2	18711366
rs1374910	3	185813873	T	IGF2BP2	21573907
rs16861329	3	186948673	C	ST64GAL1	24509480
rs6808574	3	188022735	C	LPP	24509480
rs6815464	4	1316113	C	MAEA	22158537
rs4458523	4	6288259	G	WFS1	24509480
rs1801214	4	6301295	T	WFS1	20581827
rs7659604	4	121744359	T	NR	17554300
rs6813195	4	152599323	C	TMEM154	24509480
rs702634	5	53975590	A	ARL15	24509480
rs10461617	5	56808481	A	MAP3K1	23209189
rs4457053	5	77129124	G	ZBED3	20581827
rs319598	5	134904545	C	PCBD2	24509480
rs17053082	5	155967220	T	intergenic	23300278
rs9295474	6	20652486	G	CDKAL1	21490949
rs4712523	6	20657333	G	CDKAL1	19401414
rs4712524	6	20657634	G	CDKAL1	18711366
rs10946398	6	20660803	C	CDKAL1	17463249
rs7754840	6	20661019	C	CDKAL1	17463246
rs7756992	6	20679478	G	CDKAL1	17460697
rs10440833	6	20687890	A	CDKAL1	20581827
rs6931514	6	20703721	G	CDKAL1	18372903
rs9465871	6	20717024	C	CDKAL1	17554300
rs2244020	6	31379674	G	HLA-B	25102180
rs3916765	6	32717773	A	HLA-DQA2	22693455
rs9470794	6	38139068	C	ZFAND3	22158537
rs1535500	6	39316274	T	KCNK16	22158537
rs9472138	6	43844025	T	VEGFA	18372903
rs1048886	6	70579486	G	C6orf57	21490949
rs4273712	6	126643364	G	C6orf173	24509480
rs6937795	6	136970143	A	IL20RA	24509480
rs642858	6	139952510	A	intergenic	21490949
rs7795991	7	13861106	G	ETV1	24509480
rs17168486	7	14858657	T	DGKB	24509480
rs864745	7	28140937	T	JAZF1	18372903

Table 3 continued

rsID[1]	Chr[2]	Pos[2]	Risk Allele[3]	Gene(s)[4]	PubMed ID[5]
rs849134	7	28156603	A	JAZF1	20581827
rs849135	7	28156794	G	JAZF1	24509480
rs10231619	7	43280995	T	HECW1	25102180
rs7636	7	100892456	A	ACHE	21490949
rs6467136	7	127524904	G	PAX4, GCC1	22158537
rs10229583	7	127606849	G	ARF5, PAX4, SND1	23532257
rs791595	7	128222749	A	MIR129, LEP	23945395
rs972283	7	130782095	G	KLF14	20581827
rs516946	8	41661730	C	ANK1	24509480
rs7003257	8	67701155	T	CPA6	25102180
rs17359493	8	94844683	G	INTS8	25102180
rs7845219	8	94925274	T	TP53INP1	24509480
rs896854	8	94948283	T	TP53INP1	20581827
rs13266634	8	117172544	C	SLC30A8	17293876
rs3802177	8	117172786	G	SLC30A8	20581827
rs1561927	8	128555832	C	TMEM75	24509480
rs4527850	8	133184606	T	WISP1	23300278
rs5219	9	22029548	T	KCNJ11	17463246
rs2383208	9	22132077	A	CDKN2A, CDKN2B	19401414
rs11257655	10	12265895	T	CDC123	22961080
rs10906115	10	12272998	A	CDC123, CAMK1D	20862305
rs12779790	10	12286011	G	CDC123, CAMK1D	18372903
rs2812533	10	69692529	C	C10orf35	24509480
rs12571751	10	79182874	A	ZMIZ1	24509480
rs10788575	10	88008827	A	PTEN	24509480
rs6583826	10	92588073	G	KIF11	21490949
rs1111875	10	92703125	C	HHEX	17463246
rs5015480	10	92705802	C	HHEX	17463249
rs34872471	10	112994312	C	TCF7L2	25483131
rs7901695	10	112994329	C	TCF7L2	17463249
rs4506565	10	112996282	T	TCF7L2	17554300
rs10886471	10	119389891	C	GRK5	22961080
rs10510110	10	122432914	C	PLEKHA1	24509480
rs10741243	10	131149699	G	TCERG1L	21490949
rs3842770	11	2157440	A	INS-IGF2	25102180
rs11043007	11	2183058	G	ASCL2, TH	25102180
rs231362	11	2670241	G	KCNQ1	20581827
rs231356	11	2684113	T	KCNQ1	25102180
rs2237892	11	2818521	C	KCNQ1	18711367
rs163182	11	2822986	C	KCNQ1	21799836

Table 3 continued

rsID[1]	Chr[2]	Pos[2]	Risk Allele[3]	Gene(s)[4]	PubMed ID[5]
rs163184	11	2825839	G	KCNQ1	24509480
rs2283228	11	2828300	A	KCNQ1	25102180
rs2237895	11	2835964	C	KCNQ1	20174558
rs2237897	11	2837316	C	KCNQ1	18711366
rs2722769	11	11206827	C	GALNTL4, LOC729013	22238593
rs5215	11	17387083	C	KCNJ11	17463249
rs9300039	11	41893816	C	intergenic	17463248
rs1552224	11	72722053	A	CENTD2	20581827
rs1387153	11	92940662	T	MTNR1B	20581827
rs10830963	11	92975544	G	MTNR1B	24509480
rs7107217	11	129603795	C	TMEM45B, BARX2	22238593
rs10842994	12	27812217	C	KLHDC5	24509480
rs12304921	12	50963759	G	NR	17554300
rs1153188	12	54705212	A	DCD	18372903
rs1531343	12	65781114	C	HMGA2	20581827
rs2261181	12	65818538	T	HMGA2	24509480
rs343092	12	65857160	T	HMGA2	25102180
rs1495377	12	71183321	G	NR	17554300
rs4760790	12	71241014	A	LGR5, TSPAN8	20581827
rs7961581	12	71269322	C	LGR5, TSPAN8	18372903
rs7305618	12	120965129	C	HNF1A	21573907
rs12427353	12	120989098	G	HNF1A	24509480
rs7957197	12	121022883	T	HNF1A	20581827
rs1727313	12	123156306	C	MPHOSPH9	24509480
rs10507349	13	26207391	G	RNF6	24509480
rs1359790	13	80143021	G	SPRY2	20862305
rs730570	14	100676553	G	C14orf70	21573907
rs7403531	15	38530704	T	RASGRP1	22961080
rs335810	15	60076007	A	ANXA2	25102180
rs7163757	15	62099409	C	C2CD4A	24509480
rs7172432	15	62104190	A	C2CD4A	20818381
rs1436955	15	62112183	C	C2CD4B	20862305
rs7178572	15	77454848	G	HMG20A	22693455
rs7119	15	77485290	T	HMG20A	21490949
rs11634397	15	80139880	G	ZFAND6	20581827
rs2028299	15	89831025	C	AP3S2	21874001
rs8042680	15	90978107	A	PRC1	20581827
rs12899811	15	91000846	G	PRC1	24509480
rs8050136	16	53782363	A	FTO	17463248
rs9936385	16	53785257	C	FTO	24509480

Table 3 continued

rsID[1]	Chr[2]	Pos[2]	Risk Allele[3]	Gene(s)[4]	PubMed ID[5]
rs9939609	16	53786615	A	FTO	17554300
rs11642841	16	53811575	A	FTO	20581827
rs17797882	16	79373021	T	WWOX	22158537
rs623323	17	796780	T	NXN	23300278
rs312457	17	7037074	G	SLC16A13	23945395
rs4430796	17	37738049	G	HNF1B, TCF2	20581827
rs10460009	18	2948031	C	LPIN2	21490949
rs8090011	18	7068463	G	LAMA1	22693455
rs275856	18	24259188	T	OSBPL1A	25483131
rs12970134	18	60217517	A	MC4R	24509480
rs3786897	19	33402102	A	PEPD	22158537
rs6017317	20	44318326	G	FITM2, R3HDML,	22158537
rs4812829	20	44360627	A	HNF4A	21874001
rs328506	20	57454548	C	CTCFL, RBM38,	23300278
rs2833610	21	32012873	A	HUNK	21490949

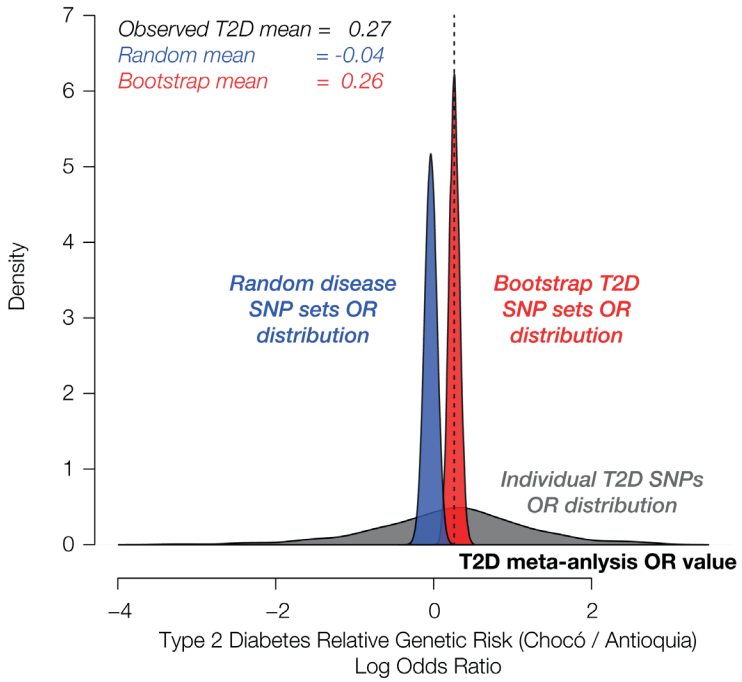

Figure 22. Distributions of T2D SNP OR values along with control analysis distributions.

The distribution of OR values for the 165 individual T2D-associated SNPs is shown in gray, and the observed T2D meta-analysis OR value is indicated with a dashed line. The bootstrap T2D SNP set OR value distribution is shown in red, and the random disease-associated SNP set OR value distribution is shown in blue. Mean values for all three distributions are shown.

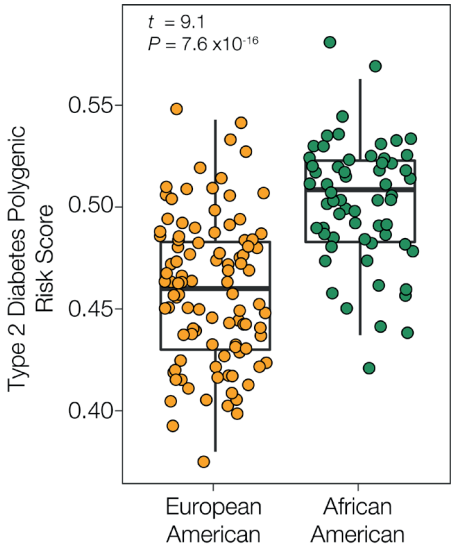

Figure 23. Type 2 diabetes polygenic risk score distributions for European-American (orange) and African-American (green) populations from the US.

The significance of the difference between the two distributions, based on the Student's t-test, is shown.

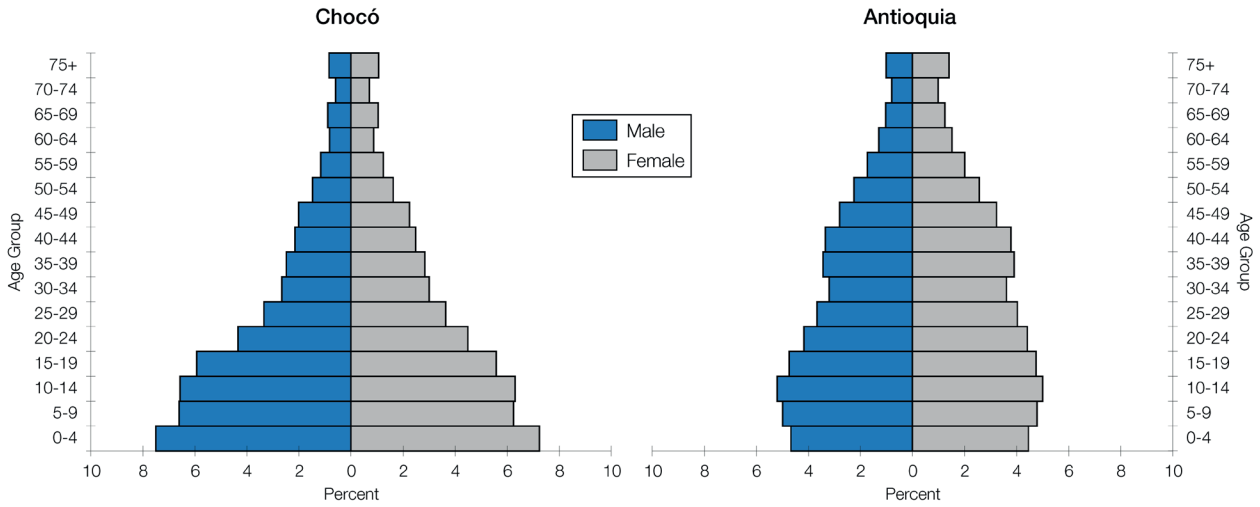

Figure 24. Age pyramids for Chocó and Antioquia.

The percentages of males and females in each population are shown for different age groups.

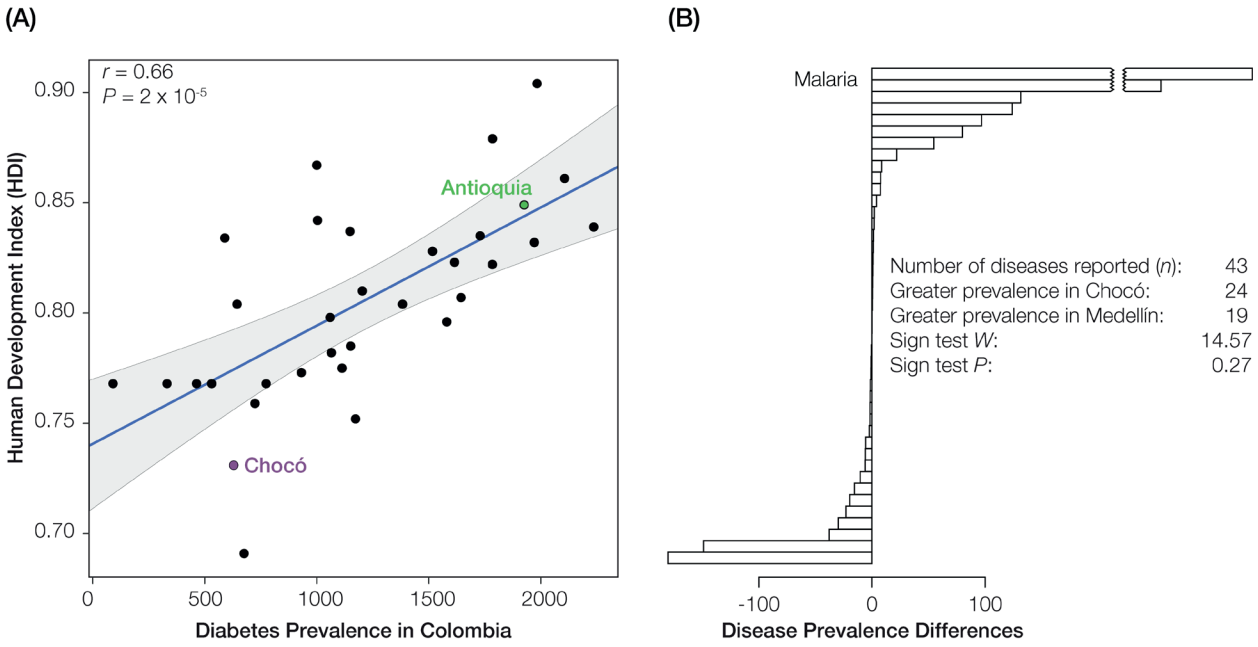

Figure 25. Economic development and disease prevalence reporting in Chocó and Antioquia.

(A) Regression of the human development index (HDI, y-axis) against diabetes prevalence estimates (x-axis) for Colombian administrative departments (i.e., states). The linear trend line is shown in blue with 95% CI in gray. The values of r and P for the Pearson correlation coefficient of the regression are shown. (B) Disease prevalence estimate differences (Chocó – Antioquia) for 43 reported diseases. The values of W and P for a binomial sign test of consistent differences between pairs of disease prevalence estimates are shown. Note that the x-axis breaks at the two highest values for malaria, which has a far higher prevalence in Chocó.

.

CHAPTER 5. ANCESTRY EFFECTS ON DIABETES GENETIC RISK INFERENCE IN HISPANIC/LATINO POPULATIONS

5.1 Abstract

5.1.1 Background

Hispanic/Latino (HL) populations bear a disproportionately high burden of type 2 diabetes (T2D). The ability to predict T2D genetic risk using polygenic risk scores (PRS) offers great promise for improved screening and prevention. However, there are a number of complications related to the accurate inference of genetic risk across HL populations with distinct ancestry profiles.

5.1.2 Results

We investigated how ancestry affects the inference of T2D genetic risk using PRS in diverse HL populations from Colombia and the United States (US). In Colombia, we compared T2D genetic risk for the Mestizo population of Antioquia to the Afro-Colombian population of Chocó. In the US, we compared European-American versus Mexican-American populations. T2D genetic risk in these HL populations is positively correlated with African and Native American ancestry and negatively correlated with European ancestry. The inferred relative risk of T2D is robust to differences in the ancestry of the cohorts used for variant discovery. Nevertheless, explicit consideration of genetic ancestry can yield more reliable cross-population genetic risk inferences.

5.1.3 Conclusions

In particular, T2D associations that replicate across populations provide for more reliable risk inference, and modeling population-specific frequencies of ancestral and derived risk alleles can help control for biases in PRS estimation.

5.2 Background

Diabetes mellitus is a global pandemic [171-173]. The prevalence of adult-onset (type 2) diabetes has nearly doubled over the last thirty years, and the number of cases has increased by more than 300 million. This increase has been driven largely by modernization and the accompanying changes in diet and lifestyle. According to the International Diabetes Federation (IDF) Atlas [174], 425 million adults worldwide are currently living with diabetes, with half of them remaining undiagnosed. In the United States (US) alone, more than 100 million adults have either prediabetes or diabetes. US Hispanic/Latino (HL) populations bear a disproportionate burden of type 2 diabetes (T2D), with a prevalence almost twice as high as that of non-Hispanic whites [175, 176]. Globally, countries from the Latin America and Caribbean region show the highest diabetes prevalence compared to six other regions.

T2D is a multifactorial disease with a complex set of interacting environmental and genetic causes contributing to its etiology. Historically, risk management for T2D has been focused squarely on environmental factors, with an emphasis on changes in diet and lifestyle. Physicians have been taught to evaluate a suite of clinically measurable risk factors, e.g., weight and blood pressure along with blood sugar and cholesterol levels, in assessing patients' likelihood of developing T2D. In addition to these clinical features, family history and race/ethnicity are also widely recognized as T2D risk factors,

underscoring genetic contributions to disease expression. Indeed, genetic factors have been estimated to account for 20-80% of the variance in T2D development [177-179]. It follows that an understanding of individual patients' genetic risk should become part of the standard of care for T2D screening and prevention.

Individuals' risk for common heritable diseases, such as T2D, can be quantified as polygenic risk scores (PRS) [180]. The ability to calculate PRS rests on genome-wide association studies (GWAS), which characterize specific genetic variants (alleles) that increase disease risk [181]. GWAS typically uncover numerous variants across the genome, each of which contributes a small fraction of the overall disease risk. PRS can be computed by summing the number of risk-increasing alleles in individuals' genomes, and scores can be weighted by the effect sizes of the risk alleles [182]. This approach to inferring genetic risk works very well when it is applied to patient cohorts from the same populations where the GWAS was conducted. However, the extent to which genetic risk can be accurately calculated across populations with divergent ancestries is a matter of contention [120, 183]. On the one hand, many GWAS are highly replicable, with the same variants often discovered in multiple populations [22, 184]. On the other hand, recent studies have shown that differences in genetic ancestry can lead to misestimation of PRS across populations [23, 29, 185].

The challenge of accurate PRS estimation across ancestry groups is particularly pressing for HL populations. First, there is a severe bias towards European ancestry cohorts in GWAS. As of 2006, only 0.06% of GWAS samples were from HL cohorts, and the fraction had only risen slightly to 0.54% by 2016 [16, 32]. Second, HL is a politically inspired, pan-ethnic label that does not correspond to any natural (i.e., genetic)

classification of human populations [186]. Individuals with origins in Latin America typically have three-way ancestry contributions from African, European, and Native American source populations, and they can differ dramatically with respect to the relative proportions of each [75-78, 187]. Even neighboring populations from within the same Latin American country can show widely divergent ancestry profiles [188]. Accordingly, the extent to which existing GWAS variants can be used to infer genetic risk among diverse HL populations accurately is currently unknown.

In this study, we explored the relationship between ancestry and T2D genetic risk inference in HL populations from Colombia and the US. We found that T2D genetic risk is positively correlated with African and Native American ancestry and negatively correlated with European ancestry, consistent with epidemiological results. We also show that T2D genetic risk inference holds up well across different GWAS ancestry cohorts and propose an approach whereby ancestry information can be used to support cross-population risk inference.

5.3 Materials and Methods

5.3.1 Diabetes epidemiological data

Data on the worldwide prevalence of diabetes mellitus were taken from The World Bank [189]. Worldwide diabetes prevalence values are expressed as the percentage of the population between the ages of 20 and 79 diagnosed with diabetes. Prevalence values are reported for 264 countries, which were broken down into seven World Health Organization (WHO) regions and four WHO income groups. Data on the prevalence of diabetes for the United States (US) were taken from the American Diabetes Association [190]. US diabetes

prevalence values are expressed as the age-adjusted percentage of the population diagnosed with diabetes. Prevalence values are broken down by the US census self-identified race/ethnicity groups and further sub-divided into country/region of origin for individuals who self-identify as Hispanic/Latino (HL). Diabetes prevalence values for the European-American (EA) and Mexican-American (MA) populations were taken from the Utah Department of Public Health [191] and the County of Los Angeles Public Health agency [192]. Note that these US diabetes prevalence values correspond to the specific populations sampled as part of the 1000 Genomes Project and used for genetic risk inference (see Methods section on Type 2 diabetes (T2D) genetic risk inference).

5.3.2 Genome-wide association study (GWAS) data

GWAS data were taken from the NHGRI-EBI GWAS Catalog [181]. All reported GWAS (as of 3/31/2018) were characterized with respect to the trait under consideration and the ancestry of the study cohort. GWAS cohorts were characterized as African, East Asian, European, Hispanic/Latino, or Native American following the GWAS Catalog framework for the representation of ancestry data in genomic studies [193]. The total number of single-nucleotide polymorphism (SNP) associations that reach the GWAS Catalog significance threshold ($P<1x10^{-5}$) were recorded for each GWAS trait. For each T2D SNP association, we recorded the study(ies) where it was reported, the cohort ancestry, the SNP identifier, its chromosomal location, and the identity of the trait-increasing effect allele. T2D GWAS summary statistics for a trans-ethnic meta-analysis, which integrated cohorts with four distinct ancestries, were taken from the DIAGRAM consortium http://diagram-consortium.org/ [194, 195]. For these data, the GWAS SNP

effect alleles, ancestry-specific directions of effect, effect sizes, and *P*-values were recorded.

5.3.3 Type 2 diabetes (T2D) genetic risk inference

Whole genome sequences from the 1000 Genomes Project [196] and imputed whole genome genotypes from the ChocoGen Research Project https://www.chocogen.com [197] were used for T2D polygenic risk score (PRS) calculation (Table 3). For the 1000 Genomes Project data, SNP data were taken from the phase 3 data release for one Colombian population – Colombians from Medellín, Colombia – and two US populations: Utah Residents (CEPH) with Northern and Western European Ancestry and Mexican Ancestry from Los Angeles USA. The 1000 Genomes Project human genome sequence data are de-identified and made publicly available for research use without restriction. For the ChocoGen Research Project, whole genome genotypes for sample donors were characterized using the Illumina HumanOmniExpress-24 SNP array as previously described, yielding ~500,000 SNPs per individual [188, 197]. The genotypes were imputed using the program IMPUTE2 [198] with the 1000 Genomes Project phase 3 haplotype reference panel [199] as previously described [200], yielding ~35 million additional SNPs across all samples. The ChocoGen project was conducted with the approval of the Ethics Committee of the Universidad Tecnológica del Chocó (ACTA Nº 01-v1), and all sample donors signed informed consent documents.

Table 4. Populations analyzed in this study.

[a]1KGP = 1000 Genomes Project. [b]n = number of sample donors per population

Data Source[a]	Population Description	Population Name	n^b
ChocoGen	Chocoano in Quibdó, Colombia	Chocó	94
1KGP	Colombian in Medellin, Colombia	Antioquia	94
1KGP	Yoruba in Ibadan, Nigeria	African	108
1KGP	Iberian populations in Spain	European	107
1KGP	Utah residents with NW European ancestry	European-American (EA)	99
1KGP	Mexican Ancestry from Los Angeles USA	Mexican-American (MA)	64
1KGP	Peruvian in Lima, Peru	Native American	85

For each genome, an unweighted T2D PRS was computed by calculating the normalized sum of the number of T2D SNP effect alleles found in the genome [182]. It should be noted that T2D PRS were not weighted by SNP effect sizes owing to the fact that the T2D SNP associations used here were curated from multiple studies whose effect sizes cannot be accurately combined [184]. T2D PRS were calculated as in Equation (1 T2D PRS were compared to individuals' continental genetic ancestry fractions – African, European, and Native American – which were taken from our previous studies [188, 200].

T2D PRS were computed for the Colombian and US populations using an unpruned set of 165 T2D-associated SNPs along with a reduced linkage disequilibrium (LD) pruned set of 42 SNPs. LD pruning was performed on the four Colombian, and US populations

analyzed here using the program PLINK [201] with 2000 SNP window size and a threshold of $r^2 > 0.1$, where r^2 corresponds to the level of linkage disequilibrium between pairs of SNPs in the window. An additional round of LD clumping was performed on the DIAGRAM GWAS summary statistic data using the LDpred program, with the same suggested window size of 2000 SNPs [202]. LDpred uses the LDscore method to choose the highest effect size SNP for each LD window and subsequently reweights the effect sizes for all retained SNPs.

5.3.4 Genetic ancestry and T2D risk

The program ADMIXTURE was used to compute the three-way continental ancestry percentages – African, European, and Native American – for all individuals from the Colombian and US populations analyzed here [203]. The modern Colombian and US populations were compared to the proxy ancestral reference populations shown in Table 4, with ADMIXTURE run for K=3 ancestral components, corresponding to each of the three continental population groups that admixed to form modern American populations. This process yields a vector of three ancestry fractions for any individual admixed genome sampled from the modern populations: $f_{African}$, $f_{European}$, $f_{NativeAmerican}$ (Additional file 2: Figure S1). Then, for each of the three continental ancestry components, individuals' continental ancestry fractions were regressed against their T2D PRS using unweighted ordinary least squares regression (OLS) with the lm function in R (Equation (**5**). The resulting OLS produces: β_0, the model β or slope; the standard error of the model; the r^2 value describing the model's fit; the model t-statistic; and a two-tailed *P*-value. For visualization purposes, a best fit line with confidence intervals was computed using local polynomial regression (loess).

5.4 Results

5.4.1 Diabetes prevalence and population disparities

Diabetes is characterized by an extremely high disease burden along with pronounced disparities in prevalence among countries, regions, and income groups worldwide (Figure 26A and B). It should be noted that, while these prevalence data are not broken down into diabetes types, the vast majority of diabetes cases correspond to adult-onset, non-insulin-dependent, type 2 diabetes (T2D). The US is no exception to this trend; there is a high overall diabetes prevalence in the country and marked disparities among racial and ethnic groups (Figure 26C). Native Americans, African Americans, and Hispanic/Latino (HL) populations bear a disproportionately high share of the diabetes disease burden in the US compared to Asian Americans and European Americans. Interestingly, there are also notable disparities within ethnic groups. HL populations with distinct origins in Latin America can have very different diabetes prevalence (Figure 26D). Individuals from South America show diabetes prevalence close to what is seen for Asian Americans, whereas Mexican Americans show a two-times greater prevalence, close to what is seen for Native Americans. Among HL regional groups, diabetes prevalence can also differ between males and females in a group-specific manner.

The observed diabetes prevalence disparities among HL groups with distinct origins begs an explanation. Diabetes is a complex common disease with multifactorial causes, including genetic and environmental effects, along with interactions between them. Nevertheless, T2D, in particular, is strongly genetically influenced with estimates of heritability ranging from 20 to 80% [177-179]. Furthermore, genetic ancestry is known to

impact the burden T2D; both African and Native American ancestry have been associated with increased T2D prevalence [128, 130, 204-206]. Thus, one may naively expect to observe more uniformity in T2D prevalence within a single ethnic group. However, the pan-ethnic HL label does not, in fact, correspond to a 'natural' group with shared genetic ancestry. Rather, HL groups encompass an extraordinarily diverse set of populations, which are characterized by distinct combinations of ancestry from Africa, Europe, and the Americas [75-78, 187]. Additionally, the Native American component of HL ancestry varies substantially according to the regional origins of the populations [188, 207, 208]. With this in mind, we have been investigating the contributions of ancestry to genetic risk and T2D health disparities in diverse HL populations.

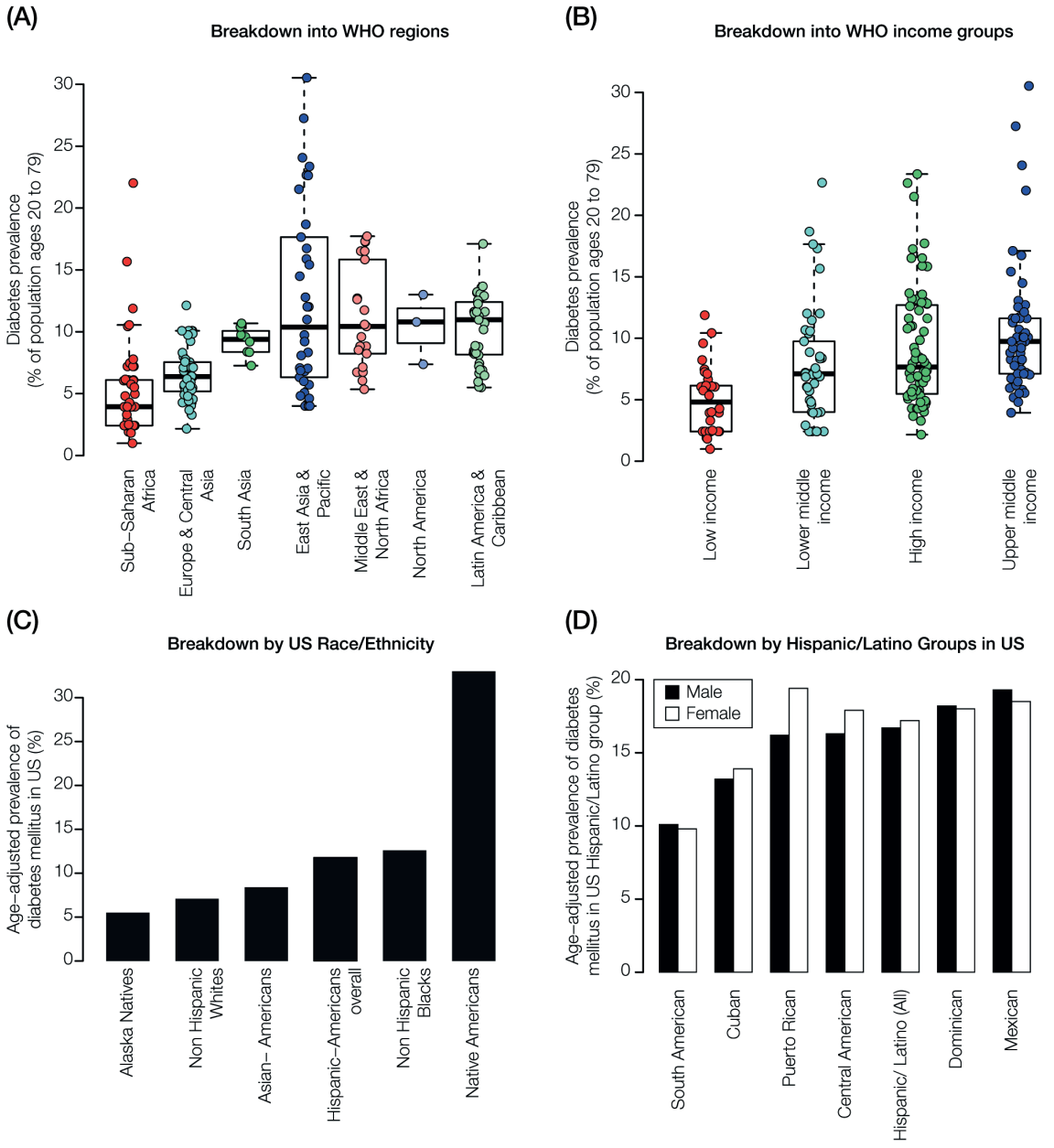

Figure 26. Diabetes global prevalence and population disparities.

(A) Diabetes prevalence distributions are shown for (A) the seven world health organization (WHO) geographic regions and (B) the four WHO income groups. (C) Diabetes prevalence for the United States (US) census race/ethnicity groups. (D) Hispanic/Latino (HL) diabetes prevalence in the US broken down by country (region) of origin and shown separately for males (black) and females (white).

5.4.2 GWAS ancestry bias and T2D risk inference

The power to infer genetic risk for complex common diseases, such as T2D, has exploded in recent years owing to the accumulation of GWAS for a wide variety of health-related traits [180, 181]. GWAS yield lists of trait SNP associations, including the identity of trait-increasing effect alleles, each of which slightly increases the risk of disease. Accordingly, an individual's genetic risk for a given trait can be estimated as a polygenic risk score (PRS), which is calculated as the normalized sum of risk (effect) alleles encoded in their genome. However, the overwhelming bias towards European cohorts in GWAS [16, 32] presents a major challenge to this paradigm. Specifically, the extent to which PRS can be accurately inferred across population groups with distinct ancestry profiles is a matter of great concern [120, 183]. On the one hand, many robust SNP associations are known to replicate across populations [22, 184]. On the other hand, GWAS SNP ascertainment biases and demographic processes have been shown to yield systematic errors in PRS calculation across populations [23, 29, 185].

Here, we aimed to explore the effects of ancestry on the calculation of PRS for T2D across diverse populations. In support of this effort, we found that T2D is distinct compared to GWAS for most other traits in several respects, largely owing to the intensity of focus on the genetic architecture of the disease and its epidemiological importance for populations across the world. T2D has the most independent studies of any trait in the NHGRI-EBI GWAS catalog (Figure 27A), and it has among the most SNP associations reported for any trait (Figure 27B). Perhaps even more importantly, for our purposes, T2D cohorts show substantially more ancestry diversity than typical GWAS traits (Figure 27C). A slight majority of T2D GWAS cohorts have European ancestry, but there are a

substantial number of cohorts with East Asian, African, and HL ancestry. A number of T2D GWAS have employed a trans-ethnic study design, whereby cohorts with distinct ancestries are combined to increase the reliability of discovered SNP associations [194, 195]. Taken together, the large number of T2D studies with diverse ancestry cohorts and the large number of T2D associations provide resolution for our efforts to (i) calculate PRS across diverse populations and (ii) assess the impact of ancestry on predicted T2D genetic risk.

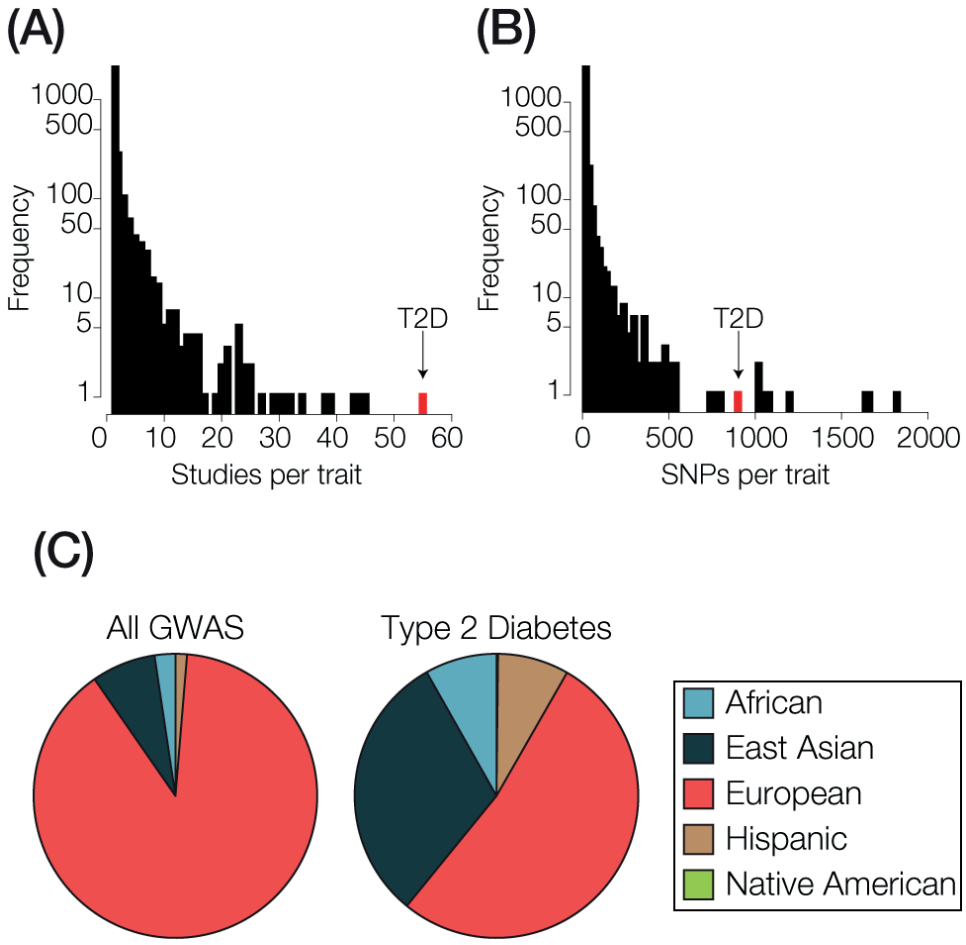

Figure 27. Genome wide association studies (GWAS) on type 2 diabetes (T2D).

The number of (A) GWAS and the number of (B) SNP-associations per GWAS trait are shown, with T2D values in red. (C) The fractions of continental ancestry groups represented in GWAS cohorts are shown for all GWAS and for T2D GWAS alone.

5.4.3 Ancestry and T2D genetic risk inference: Colombia

We first explored the relationship between ancestry and T2D genetic risk for the Colombian populations of Antioquia and Chocó. Although these two administrative departments (states) share a common border, their populations were historically isolated and showed very distinct ancestry profiles. The population of Antioquia has majority European ancestry (75%) followed by Native American (18%) and African (7%) fractions, whereas the ancestry of Chocó is primarily African (76%) with smaller European (13%) and Native American (11%) components [188]. Genome sequences were characterized for individuals from the two populations, and T2D PRS were computed for all individuals as described in the Methods. The distributions of T2D PRS for the two populations were then compared in order to assess their relative genetic risk. Consistent with previous results [200], we found that Chocó has significantly higher predicted genetic risk for T2D compared to Antioquia (Figure 28A), and the higher genetic risk for T2D in Chocó is correlated with African ancestry (Figure 28B). The elevated T2D risk for Chocó can be observed when all 165 T2D-associated SNPs are used for PRS calculation (Figure 28) or when a reduced set of 42 linkage disequilibrium (LD) pruned SNPs is used (Additional file 2: Figure S2 panels A & B). These findings are consistent with reports from the US showing a correlation between T2D genetic risk and African ancestry [132], and African Americans are known to have substantially higher T2D prevalence compared to European Americans [128, 130, 204, 206]. In Colombia, however, Antioquia shows approximately three-times higher observed T2D prevalence compared to Chocó (Figure 28C), in direct

contrast to the predicted genetic risk for the two populations and the epidemiological data from the US.

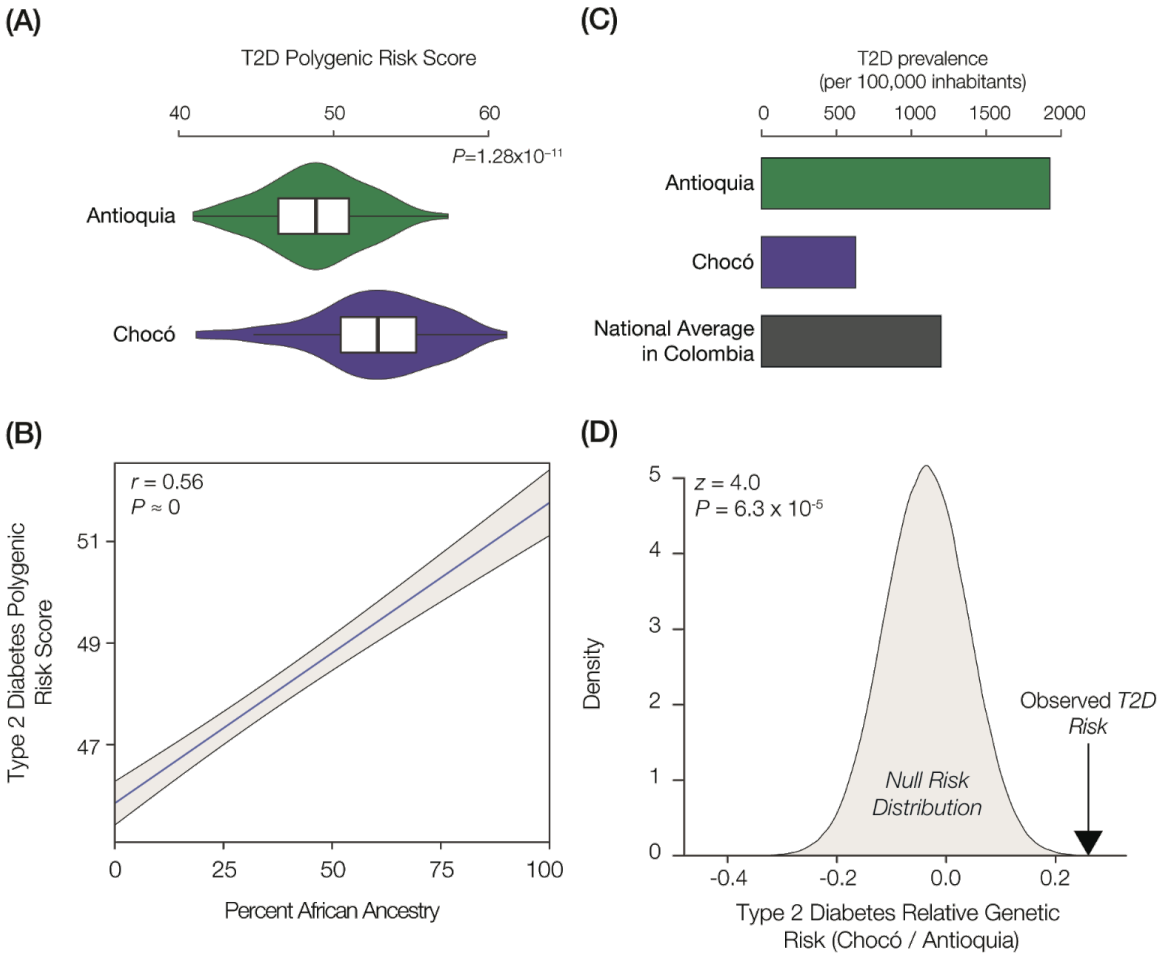

Figure 28. T2D genetic risk and observed prevalence in Colombia.

(A) T2D polygenic risk score distributions are shown for Antioquia (green) and Chocó (purple). (B) T2D polygenic risk scores for individuals from Antioquia and Chocó regressed against their percent African ancestry. (C) Observed T2D diabetes prevalence for Antioquia (green), Chocó (purple), and Colombia overall (gray). (D) Observed T2D relative genetic risk Chocó/Antioquia compared to the null distribution of relative genetic risk between the two populations.

We previously attributed the difference between the relative predicted genetic risk of T2D for the two Colombian populations and their observed T2D prevalence to gene-by-environment interactions, whereby diet and lifestyle in Chocó serve as protective factors

against T2D [200]. However, another possible explanation for this discrepancy is that there is a systematic bias in T2D PRS calculations across populations of this kind with distinct ancestry profiles [23, 29, 185]. We addressed this possibility by comparing the observed T2D relative risk for Chocó / Antioquia to a null distribution of relative risk generated by permuting 500,000 random sets of GWAS SNPs (risk alleles) of the same size as the T2D SNP set. If there were a systematic bias in the population-specific frequencies of GWAS risk alleles for the two populations, then the null distribution would be expected to show an overall increase of genetic risk in Chocó. We do not observe any such bias; the observed relative risk of T2D is significantly greater than the null expectation (Figure 28D).

As previously described, the major source of bias for cross-population PRS calculation is attributed to the vast over-representation of European cohort GWAS. It is possible that GWAS SNPs discovered in European study cohorts will not accurately capture genetic risk in non-European cohorts. This problem could be even more exacerbated in the case of the admixed Colombian populations studied here, one of which looks more European while the other is more African. The fact that T2D has been the subject of numerous GWAS across diverse population cohorts (Figure 27) provides an opportunity to interrogate this potential bias. To do so, we characterized T2D GWAS variants according to the ancestry of the study cohorts where they were discovered. We then re-calculated population-specific T2D PRS distributions for each ancestry separately. We were able to classify T2D SNPs into five different ancestry profiles, three of which showed significantly higher risk in Chocó and two of which yielded no significant difference (Figure 29). None of the comparisons showed significantly higher T2D risk in Antioquia, and all of the cohorts with ancestry most similar to the Colombian populations

(African, Multi-ethnic, and Admixed American) showed higher relative risk in Chocó. These results support the finding of higher genetic risk for T2D in Chocó, associated with African ancestry. They do not suggest that this finding can be attributed to GWAS SNP discovery bias.

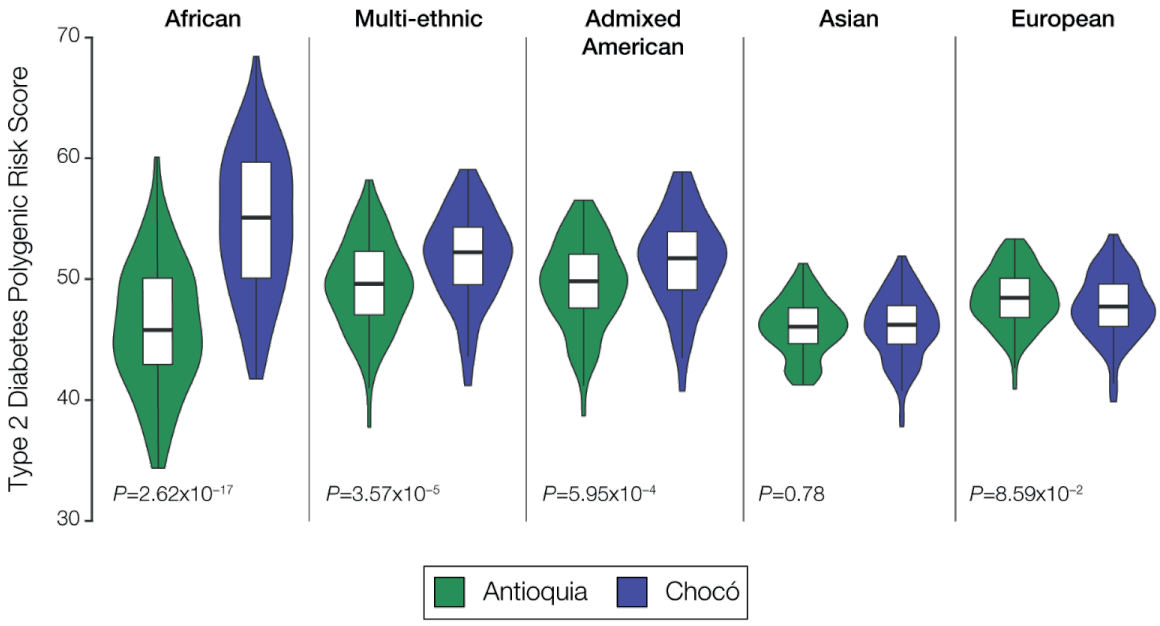

Figure 29. T2D genetic risk comparison in Colombia based on different GWAS cohort continental ancestries.

T2D polygenic risk score distributions for Antioquia (green) and Chocó (purple) are shown for SNP associations discovered in patient cohorts with distinct continental ancestries.

5.4.4 Ancestry and T2D risk inference: United States (US)

We performed a similar comparison of T2D genetic risk for European-American (EA) and Mexican-American (MA) populations in the US. With the same set of T2D SNPs used to compare genetic risk in Colombia, the MA population shows marginally higher T2D genetic risk than the EA population (Figure 30A). As was the case for Colombia, the same differences in T2D genetic risk between the US populations can be seen when all 165

T2D-associated SNPs are used for the PRS calculations (Figure 30A) or when a reduced set of 42 linkage disequilibrium (LD) pruned SNPs is used (Figure S2 panels C & D). For these two US populations, T2D genetic risk is negatively correlated with European ancestry and positively correlated with Native American ancestry (Figure 30B). However, unlike what we observed in Colombia, the relative genetic risk estimates between the two populations are consistent with the observed T2D prevalence; the MA population shows approximately two-times higher T2D prevalence than the EA population (Figure 30C).

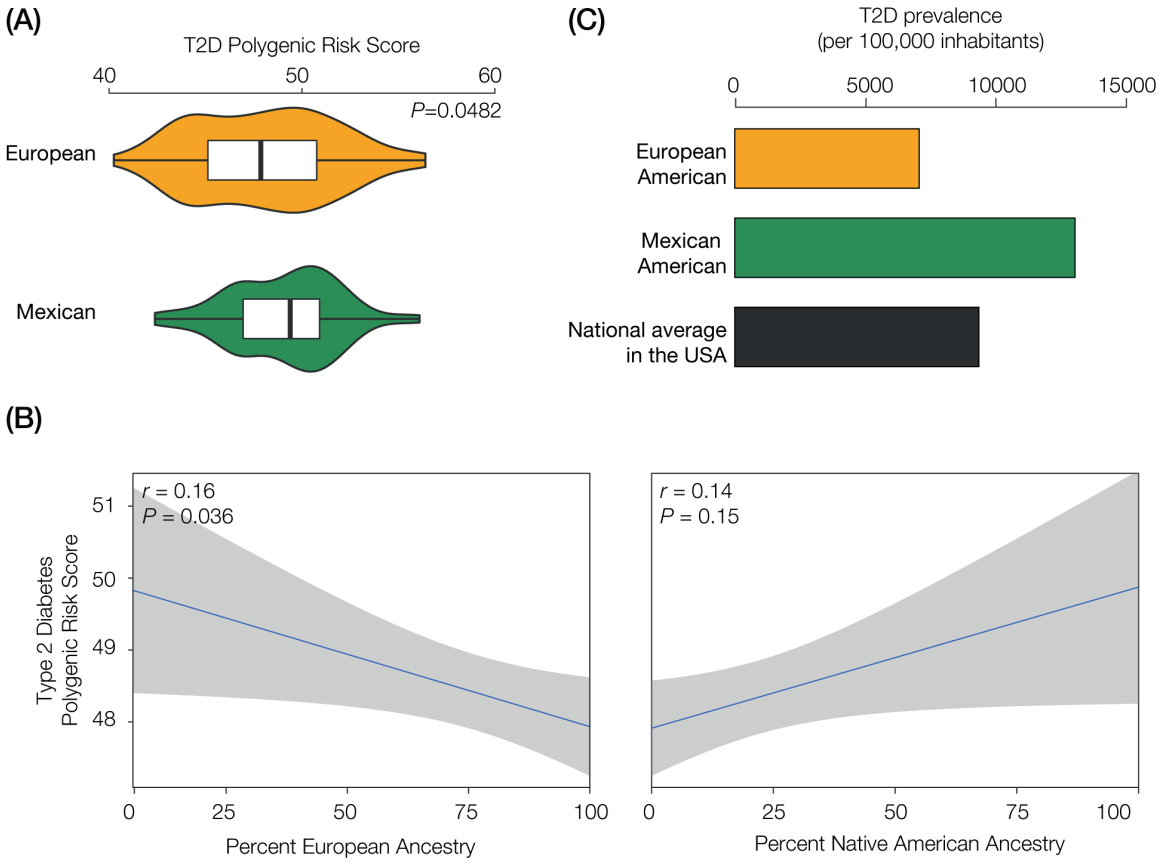

Figure 30. T2D genetic risk and observed prevalence for European-American (EA) and Mexican-American (MA) cohort populations.

(A) T2D polygenic risk score distributions are shown for EA (gold) and MA (green). (B) T2D polygenic risk scores for EA and MA individuals are regressed against their percent European and percent Native American ancestry. (C) Observed T2D diabetes prevalence values for EA (gold), MA (green), and the United States overall (gray).

Despite the consistency of the T2D genetic risk estimates and the observed prevalence values for these two populations, we wanted to explore further the contribution of genetic ancestry differences to potential biases in genetic risk calculation. To do so, we took advantage of a recent trans-ethnic GWAS meta-analysis [194, 195] to curate T2D SNPs that were discovered in one or more cohorts with distinct ancestries, including European and Mexican ancestry cohorts. We then computed T2D PRS distributions using (i) significant SNPs that showed the same direction of effect between the two ancestry cohorts, (ii) SNPs that were significant in the European ancestry cohort only, (iii) SNPs that were significant in the Mexican ancestry cohort only, and (iv) SNPs that showed different directions of ancestry-specific effects (Figure 31). The SNPs with effects that are shared between populations or effects that are population-specific all yielded higher T2D PRS in the MA compared to the EA population. The magnitude and significance of this relationship were most pronounced for the ancestry shared SNPs (Figure 31A). The SNPs with different effects between the two ancestry cohorts were the only ones that showed higher T2D PRS in the EA population (Figure 31D). These results underscore the potential utility of combining cohorts with distinct ancestries for GWAS SNP discovery, in terms of both increasing the reliability of SNP effect allele discovery and decreasing the likelihood of false discoveries. Indeed, we found that the T2D SNPs that showed shared effects across ancestry cohorts had effect size odds-ratio (OR) values almost an order of magnitude higher than SNPs with divergent ancestry-specific effects (Shared OR=2.40 versus Divergent OR=0.28).

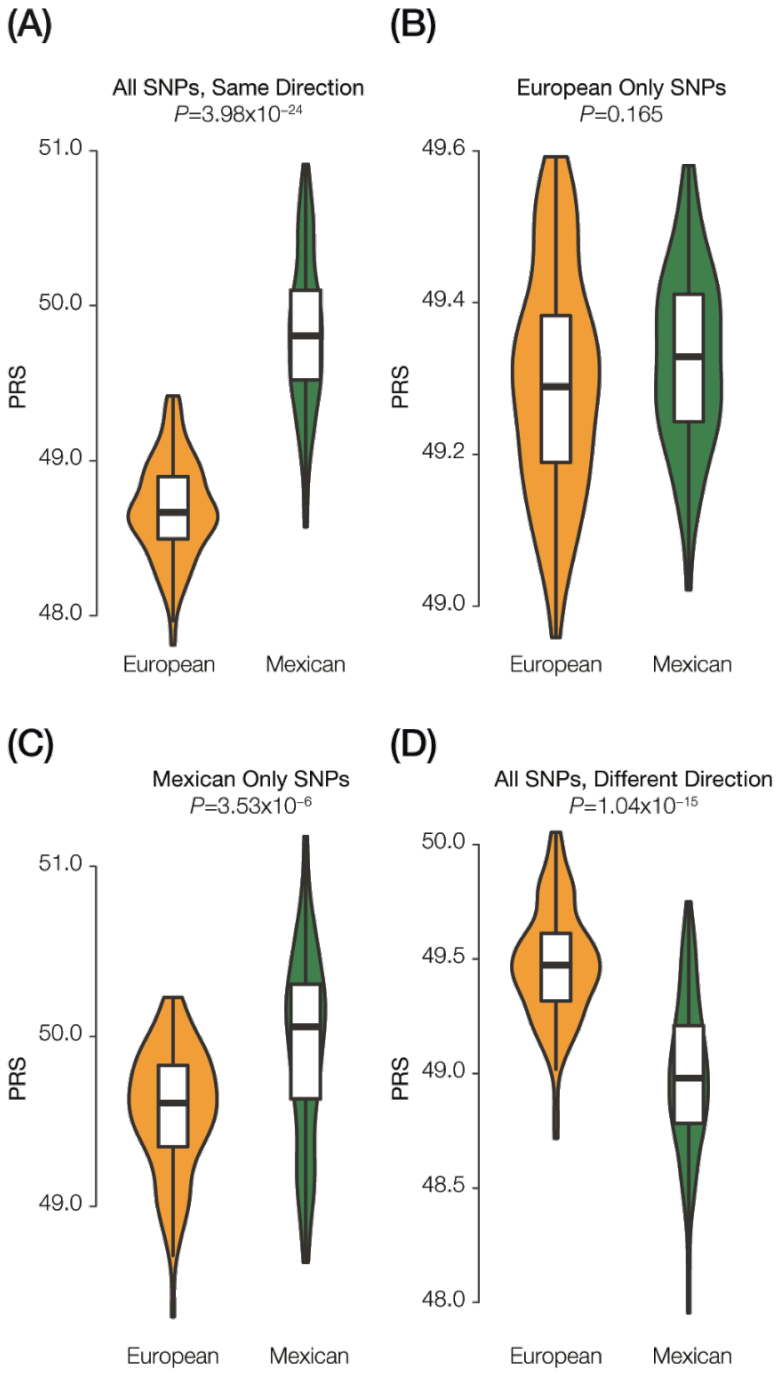

Figure 31. T2D genetic risk comparison between European-American (EA) and Mexican-American (MA) cohort populations based on ancestry-specific SNP effects.

T2D polygenic risk score distributions for EA (gold) and MA (green) populations are compared for (A) all SNPs with consistent ancestry effects, (B) SNPs with European ancestry-specific effects, (C) SNPs with Mexican ancestry-specific effects, and (D) SNPs with opposing ancestry effects.

5.4.5 Correcting for ancestry bias in T2D risk inference

A number of recent studies have underscored (i) the extreme bias of European ancestry cohorts in GWAS [16, 32] and (ii) the corollary potential to misestimate genetic risk across populations with diverse ancestries [23, 29, 120, 183, 185]. Kim et al. identified two potential sources of bias for cross-population ancestry risk inference [23], which we will call here SNP ascertainment bias and SNP discovery bias. SNP ascertainment bias is related to the fact that SNP microarrays are typically used for genotyping in GWAS, and these microarrays are designed, for the most part, to capture high minor allele frequency (MAF) SNPs in European populations. This will lead to the ascertainment of SNPs with higher MAF in European populations compared to other global populations, particularly populations from Africa that are enriched for ancestral alleles [209]. Then, systematic differences in the proportions of derived alleles, which most often correspond to the minor allele versus ancestral alleles, may lead to directional biases in the estimation of genetic risk. SNP discovery bias is related to the increased power of GWAS to detect SNPs with higher MAF. Irrespective of microarray design, the discovery of SNPs in European cohorts will yield relatively higher MAF in European populations compared to other populations, which can also lead to misestimation of genetic risk across populations with distinct ancestries.

Here, we propose a potential control for these two sources of PRS bias, based on correction for systematic differences in the proportions of ancestral versus derived alleles in populations with distinct ancestry profiles. Ancestral alleles tend to correspond to major

alleles, whereas derived alleles most often correspond to minor alleles in discovery cohort populations. While GWAS risk alleles can be more evenly distributed across ancestral (44%) versus derived (56%) alleles, differences in the frequencies of these allele classes across populations can still introduce bias in genetic risk inference [23]. The idea behind the control that we propose here is to eliminate any possible bias owing to population-specific differences in the frequencies of ancestral versus derived alleles, which are mainly attributed to demographic factors (i.e., genetic drift).

The steps in the control are shown below. Further detail regarding the execution of each step is provided in Additional file 2 (see pages 5-7).

1. Collect trait SNP set and calculate population-specific PRS values and between-population PRS differences (ΔPRS).
2. Determine the distribution of derived allele frequencies (DAF) for trait-associated SNPs in the GWAS cohort source population.
3. Randomly sample SNP sets parameterized by this DAF distribution based on the DAFs from the distinct populations being compared (thereby eliminating between-population DAF biases).
4. Calculate between-population ΔPRS for all randomly sampled SNP sets and determine the null ΔPRS distribution.
5. Compare the observed ΔPRS to the null ΔPRS distribution and compute a z-score as the ancestry-corrected ΔPRS: $corr.\Delta PRS = (obs\Delta PRS - \mu_{null\Delta PRS})/\sigma_{null\Delta PRS}$.

An example of this control can be seen for the comparison of T2D genetic risk between the EA and MA populations (Figure 32). The observed value of ΔPRS for EA-MA is -2.08, while the null ΔPRS distribution is centered around 0 with a mean value of -0.16 and a standard deviation of 1.25. Thus, there is a slight bias in PRS calculation for the two populations. Accordingly, correcting for SNP ascertainment bias does mitigate the difference in predicted risk between the two populations, with a corrected ΔPRS value of 1.54 that is marginally significant at P=0.054. Given what we know about the higher prevalence of T2D in the MA population, we may consider this correction to be accurate, in the sense that it preserves the direction of the genetic risk difference, but conservative as it dampens the observed effect.

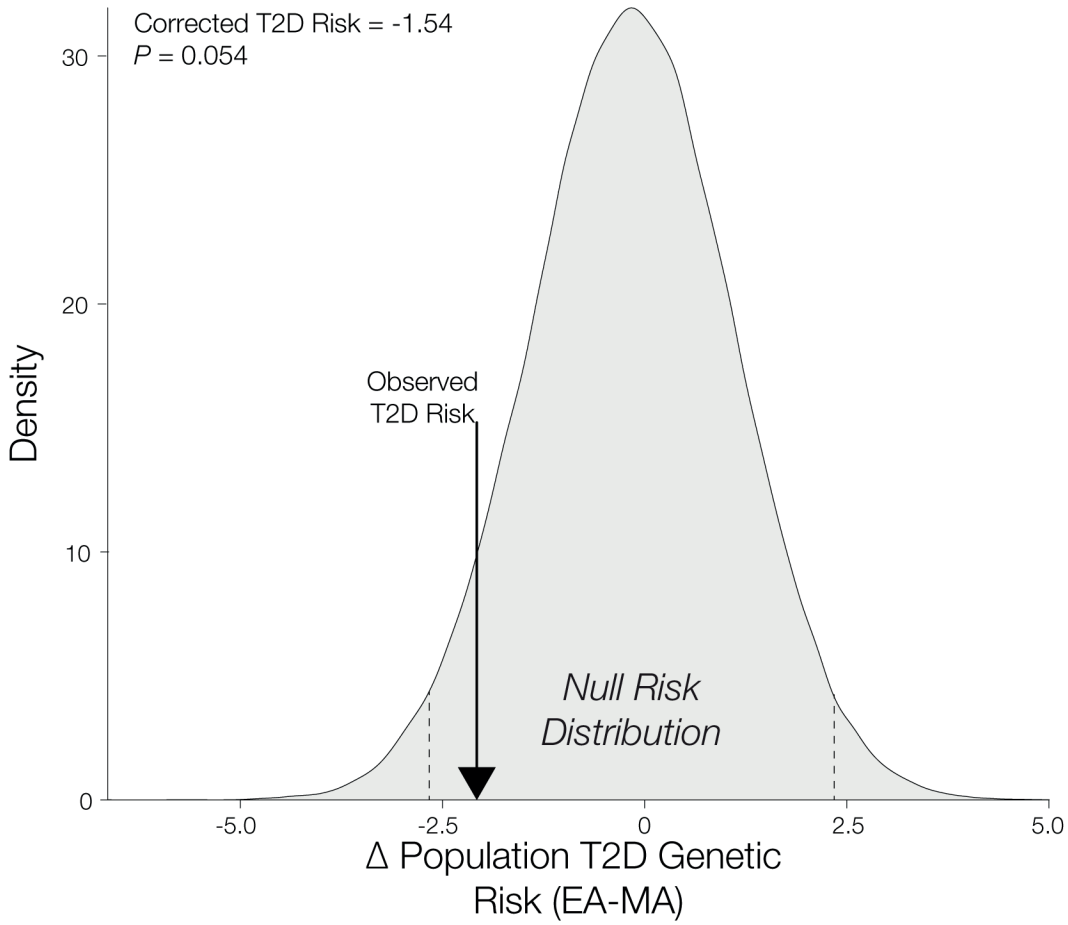

Figure 32. T2D genetic risk comparison between European-American (EA) and Mexican-American (MA) cohort populations based on ancestry-specific SNP effects.

T2D polygenic risk score distributions for EA (gold) and MA (green) populations are compared for (A) all SNPs with consistent ancestry effects, (B) SNPs with European ancestry-specific effects, (C) SNPs with Mexican ancestry-specific effects, and (D) SNPs with opposing ancestry effects.

5.5 Discussion

HL populations are burdened with a high and increasing prevalence of T2D, both in the US and in Latin America (Figure 26) [175, 176]. Recent developments in the estimation of genetic risk using PRS provide opportunities to reduce this burden through improved screening and prevention efforts [180]. Nevertheless, there are a number of challenges that need to be met in order to ensure that genetic risk of T2D, and other common heritable diseases, can be accurately predicted using PRS [120, 183]. In particular, the bias towards European ancestry cohorts in GWAS [16, 32] has the potential to limit the utility of PRS in HL populations. In addition, the extremely diverse ancestries that can be found among HL populations could lead to misestimation of genetic risk for distinct HL subgroups.

There are two broad solutions to these ancestry-related challenges to genetic risk inference: (i) more data and (ii) better methods. Obviously, more GWAS that includes cohorts that capture the genetic diversity of HL populations will go a long way towards providing the raw material, in the form of risk increasing genetic variants relevant to those same populations, which are needed to compute accurate PRS. However, given the current pace of efforts to diversify GWAS, along with the very high cost of these studies, it is unrealistic to expect the GWAS coverage of HL populations to approach that of European ancestry cohorts any time soon. In the meantime, new methods that explicitly leverage ancestry, e.g., modeling differences in allele frequencies across populations, may help to increase confidence in cross-population PRS calculation.

Here, we have shown that considering the consistency of GWAS variant effects across populations and modeling population-specific allele frequencies can increase confidence in cross-population PRS. T2D is a special case concerning common heritable diseases in the sense that it has been extensively studied via numerous GWAS, and it has the most diverse set of ancestry cohorts seen for any GWAS trait (Figure 27) [181]. In addition, recent studies have combined cohorts from different ancestries to increase confidence in the discovery of T2D associated variants [194, 195]. These facts allowed us to evaluate the extent to which GWAS variants discovered in cohorts with different ancestries yield similar PRS. The signal of T2D relative risk in Colombia is highly similar when GWAS variants discovered in different ancestry cohorts are used for PRS (Figure 29). A similar result was seen for T2D risk in the US, but in this case, consistency of T2D associations across cohorts seemed to provide more reliable PRS estimates (Figure 31). Finally, we proposed a conservative control for cross-population PRS inference based on modeling the frequencies of ancestral and derived alleles in the different populations being considered (Figure 32).

A recent study compared the utility of GWAS SNPs ascertained from EA versus HL populations for a calculating PRS in HL populations across twelve different traits [210]. While there was a wide variety of relative performance of EA SNPs across the traits, the majority of EA SNP sets showed comparable risk prediction accuracy compared to the best performing SNP sets, which included information from HL GWAS cohorts. Nevertheless, the inclusion of non-EA GWAS association results to refine the SNP weights improved accuracy across the board. The results are consistent with our findings

suggesting that information from multi-ethnic GWAS cohorts can be used to refine PRS inference.

5.6 Conclusions

One promising area for future work entails the application of machine learning methods to the inference of polygenic risk [211]. Currently, PRS calculations are based on GWAS that explicitly assume an additive model of genetic effects on traits of interest. Accordingly, standard methods for computing PRS, such as the kind we use here, entail a straightforward summation of risk alleles genome-wide. Of course, it may be more biologically realistic to assume that there are non-additive genetic effects among variants discovered by GWAS and used for PRS. If this is indeed the case, then more sophisticated machine learning algorithms may ultimately improve the accuracy of PRS calculation. The use of machine learning for polygenic risk inference is still in the very early stages; it remains to be seen if this approach will yield demonstrable improvement over current best practices.

The control we developed here for cross-population PRS inference is based on differences in ancestral versus derived allele frequencies among populations with distinct ancestry profiles. However, differences in LD across populations with divergent ancestries can also confound cross-population PRS inference. This is particularly true for African ancestry populations, which tend to have short and distinct LD blocks compared to non-African populations. Accordingly, controlling for such differences provides another promising approach for improving cross-population PRS inference. Indeed, a previous study has shown that accommodating differences in LD patterns across populations can

substantially improve the accuracy of PRS computed for distinct ancestry cohorts [212]. In the future, we plan to combine allele frequency and LD based approaches to improving the accuracy of cross-population PRS.

We employed a population-level approach to T2D genetic risk inference and evaluation in this study, comparing T2D relative genetic risk between populations to population-specific ancestry profiles and epidemiological data on observed T2D prevalence. Taken together with the robust collection of T2D variant associations from several diverse GWAS cohorts, this approach allowed us to broadly assess the impacts of ancestry on T2D genetic risk inference in HL populations. Going forward, a more rigorous assessment of PRS accuracy will require individual-level phenotype data for both model training and test sets. Data of this kind are beginning to emerge thanks to the activity of a number of diabetes research consortia along with more broadly focused biobanks that collect patient genotypes and electronic health records. We anticipate that joint analysis of individual-level genotype-phenotype data gleaned from sources of this kind will help to further develop and validate ancestry-informed approaches to T2D genetic risk inference.

5.7 Supplementary Methods

5.7.1 Genetic ancestry and admixture for Colombian and US populations

Whole genome genotypes from the HL populations in Colombia and the US were compared to genotype data from proxy source populations in Africa, Europe, and the Americas (Table 4) using the program ADMIXTURE [203] in order to calculate three-way ancestry percentages for each individual genome. The resulting ancestry percentages were then regressed against predicted T2D PRS for these same individuals, as shown in Figure

28B and Figure 30B. Details of this analysis can be found in the Materials and Methods section.

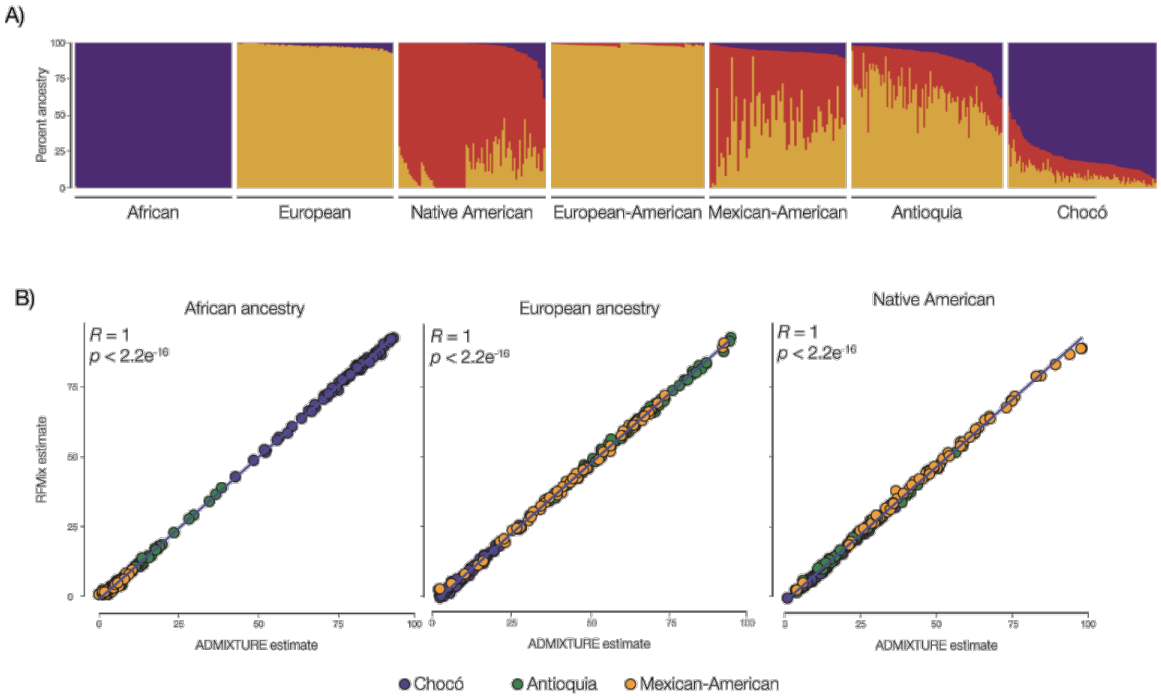

Figure 33. Ancestry and admixture patterns for the Colombian and US populations studied here.

(A) Ancestry percentages – African (blue), European (yellow), and Native American (red) – are shown for individuals sampled from proxy ancestral source populations in Africa, Europe, and the Americas along with the admixed Colombian (Antioquia and Chocó) and US populations (European- and Mexican-American). Each stacked bar represents the relative ancestry percentages for one individual from a given population. The population sources for this analysis are shown in Table 4. (B) Comparison between ADMIXTURE (x-axis) and RFMix (y-axis) continental ancestry percent estimates for the admixed populations analyzed here. Individuals from each admixed population are color-coded, as shown.

5.7.2 Effects of linkage disequilibrium (LD) on T2D genetic risk inference

Divergent patterns of LD across populations with distinct ancestry can confound cross-population PRS inference. We first controlled for the effects of LD by performing LD pruning on the initial set of 165 T2D-associated SNPs used to compute PRS in the Colombian and US populations. LD pruning was performed for all four populations together by removing variants that are linked at $r^2 > 0.1$, retaining the linked SNP with the highest minor allele frequency. Details of this analysis can be found in the Materials and Methods section. Our approach to LD pruning was intended to be very conservative (stringent) in terms of both the low r^2 threshold and the combined use of the four populations. Consistent with this intention, LD pruning reduced the total number of T2D associated SNPs for PRS analysis from 165 to 42. Nevertheless, the signal of relative T2D risk between populations remains the same for both Colombia and the US (Figure 34).

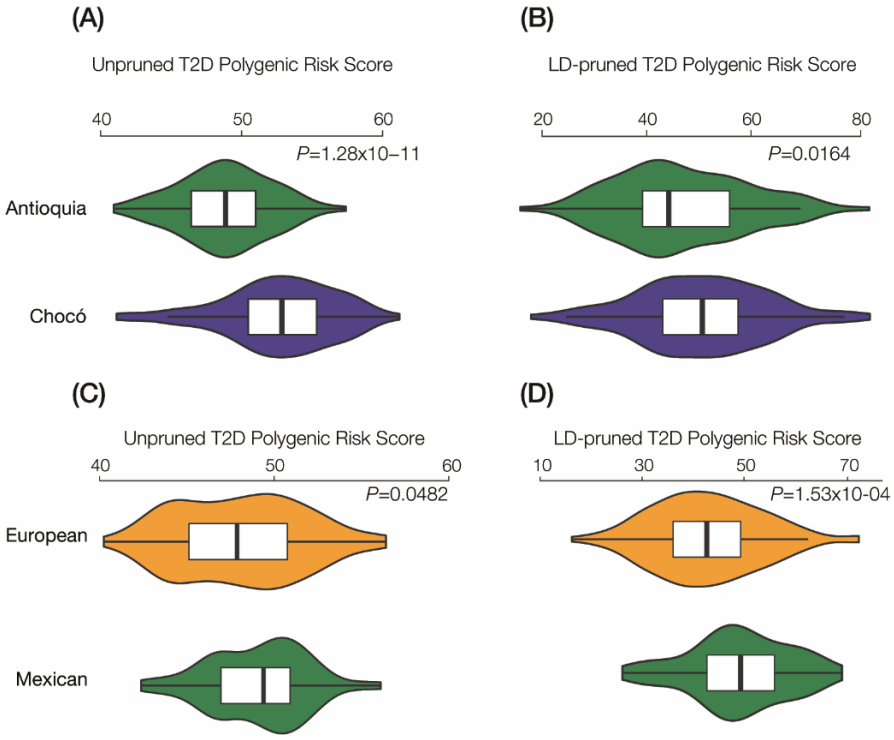

Figure 34. Effects of linkage disequilibrium (LD) pruning on T2D genetic risk.

Polygenic risk scores were calculated and presented, as shown in Figure 28 and Figure 30, using both the full unpruned set of T2D associated SNPs (n=165) and the reduced LD pruned set (n=42). Results are shown for the Colombian (panels A and B) and US populations (panels C and D).

We further attempted to control for LD differences in cross-population PRS by using the LDpred program to compute PRS for the EA and MA populations using the DIAGRAM multi-ethnic GWAS T2D SNP-association data. LDpred performs LD clumping, to correct for LD structure, along with re-weighting on SNP effect sizes, by choosing the most informative SNP within any given LD window [202]. We ran LDpred across of series of P-value thresholds (Figure 35 panels B-E) and found the results to be almost entirely consistent with the original unpruned set (Figure 35 panel A).

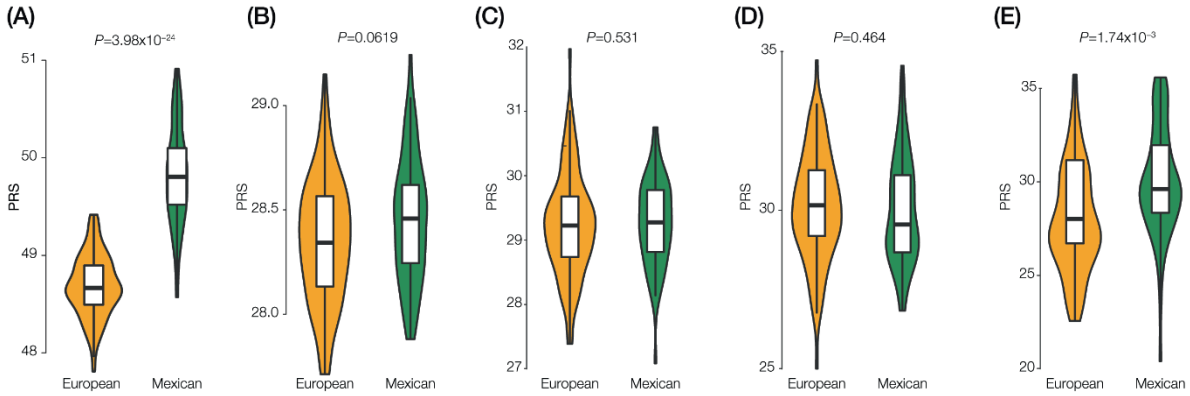

Figure 35. Effects of linkage disequilibrium (LD) clumping and *P*-value thresholding on T2D polygenic risk scores.

Distributions of population PRS using all marginally-significant variant effects from GWAS summary data (A) and from LD clumping and P-value thresholding using an LD cutoff of $r^2 > 0.1$ and P-value cutoffs of $1x10^{-1}$ (B), $1x10^{-2}$ (C), $1x10^{-3}$ (D), and $1x10^{-4}$ (E).

There are other programs available for controlling for LD when computing PRS – lassosum and PRSice-2 – but these programs use individual-level phenotype data to optimize the selection of SNPs to be included in PRS tests [213, 214]. Lassosum can be run in the pseudo validation mode without phenotypes, but doing so here on the DIAGRAM data simply yields overlapping PRS distributions between populations. Running PRSice-2 without phenotype data yields an error message and does not produce any output.

5.7.3 Correcting for ancestry bias in T2D risk inference

Our control for ancestry bias in polygenic risk inference is presented in the main text; additional details of this process are given below.

1. Collect trait SNP set and calculate population-specific PRS values and between-population PRS differences (ΔPRS).

1.1. SNP sets for traits of interest can be curated from the NHGRI-EBI GWAS Catalog at https://www.ebi.ac.uk/gwas/ or from GWAS consortium resource webpages, such as the DIAGRAM page used to curate T2D SNPs here http://diagram-consortium.org/downloads.html. SNP sets can also be curated from the literature using the NCBI PubMed database https://www.ncbi.nlm.nih.gov/pubmed/. In the case of literature searches for trait SNP sets, we recommend using the most up-to-date and comprehensive meta-analyses that can be found for any trait of interest.

1.2. PRS can be computed from whole genome genotype data by evaluating the numbers of risk alleles present in any individual genome (genotype) being analyzed. Unweighted PRS or weighted PRS can be computed. Note that unweighted PRS should be used when trait SNPs are curated from multiple studies, owing to the fact that effect sizes from different studies cannot be accurately combined.

1.3. Between-population PRS differences (ΔPRS) are computed as the difference between mean population-specific PRS values. For example, for the EA and MA populations, ΔPRS is computed as

2. Determine the distribution of derived allele frequencies (DAF) for trait-associated SNPs in the GWAS cohort source population.

2.1. PRS are computed with biallelic SNPs taken from the 1000 Genomes Project phase 3 release VCF file. This file designates each allele as ancestral or derived in the 'AA' field of the 'INFO' column.

2.2. The derived allele frequencies for (DAF) for trait-associated SNPs in the GWAS catalog are then calculated.

3. Randomly sample SNP sets parameterized by this DAF distribution based on the DAFs from the distinct populations being compared (thereby eliminating between-population DAF biases).
    3.1. The derived allele frequency (DAF) in the GWAS catalog is represented as D and the binomial probability of sampling an ancestral or derived allele is used to randomly select SNPs.
4. Calculate between-population ΔPRS for all randomly sampled SNP sets and determine the null ΔPRS distribution.
    4.1. For each simulation, population-specific PRS values for the randomly simulated SNP sets are computed as shown in #1.2 above, and the mean population-specific PRS values are then used to compute the ΔPRS value as shown in #1.3 above.
5. Compare the observed ΔPRS to the null ΔPRS distribution and compute a z-score as the ancestry- corrected ΔPRS.
    5.1. The procedure in #4 above yields a null distribution of ΔPRS values.
    5.2. The null ΔPRS value distribution from the simulated data is compared against the observed between population ΔPRS value to derive the ancestry corrected ΔPRS: $corr.\ \Delta PRS = (obs\Delta PRS - \mu_{null}\Delta PRS\ )/\sigma_{null}\Delta PRS$.

CHAPTER 6. CONCLUSIONS AND FUTURE PROSPECTS

The causes of common complex diseases remain an active area of study that bridges several scientific disciplines, and which requires researchers to disentangle the complex network of societal, environmental, and genetic factors that underlie these diseases. Through massive GWA studies, studying millions of individuals, genetics research has begun to understand some of the genetic architecture of disease. However, this effort has historically and still today been focused on individuals of European descent, contributing to increasing minority health disparities. Interrogating the genomes of diverse populations in the Americas – such as Afro-descendent populations in the US and Latin American – is a necessary step in decreasing health disparities across the board.

This dissertation work includes the development of methods that enable scientists to apply existing knowledge already derived in one population to new and diverse populations. Part of this development effort is focused on PGS, a crucial part of genomics-enabled precision medicine. In this thesis, I explored the distributions of PGS computed in diverse global populations as well as the effects of ancestry on PGS computation and accuracy using a population genetics approach in the US and Colombia. In traditional PGS, individuals are scored on the basis of SNP associations ascertained in an ancestry-matched cohort, that is, European-ancestry individuals use PGS derived from European cohort GWAS. These PGS are typically constructed using one of two methods; (1) the Top-SNP approach, in which only SNP associations that reach genome-wide significance ($p < 5 \times 10^{-8}$) are considered, and (2) the Clump and Threshold approach, where

investigators iteratively select the combination of LD-clumping r^2 and p-value threshold that produces the best score (typically measured as highest AUC).

As discussed throughout, I use a combination of these two techniques, along with several novel controls, to develop and compute PGS for complex common diseases in the US and Colombia. Differences in population PGS distributions are generally an accurate indicator of relative disparities between populations in a country, at least for Colombia. In cases where predictions do not match actual disparities, we note that there are significant socioeconomic and environmental effects that mediate the genetic component of risk. In the case of type 2 diabetes, low SES, and its association with a low-fat/high vegetable diet and increased activity levels were found to be protective against disease. Similarly, while Chocóanos were predicted to have a much lower risk for all *P. falciparum* caused malaria, the high density of parasite and mosquitoes in Chocó compared to Antioquia leads to higher rates in Chocó.

Ultimately, increased study of population genetics for diverse ancestry groups is needed to close the gap in health disparities. The major limitations of the work presented in this thesis are the lack of individual-level phenotype data and the small sample size. Unfortunately, the collection of larger datasets, which include individual-level phenotypes, is not easily accomplished from across the globe. Governments and funding agencies must invest in local efforts to study diversity.

Efforts are underway globally to expand GWA studies to include more diverse individuals. Recruiting, enrolling, and phenotyping study participants scales poorly at the individual investigator level. Thus, large scale and multi-ethnic GWAS are costly and

difficult to conduct for individual investigators and require multi-institute collaborations. Nevertheless, progress on this front has been slow but steady, resulting in multi-ethnic GWAS of hundreds of thousands of diverse individuals. Analysis efforts on these cohorts are hampered by lack of deep phenotyping – that is, phenotyping across the entire spectrum of quantitative measures and environmental exposures – which limits the new association and meta-analysis studies that can be conducted.

Researchers prefer, instead, to work with biobank-sized cohorts such as the UK Biobank, and remove diverse individuals to create more homogeneous cohorts. Thus, while the scientific community, particularly funding agencies, are beginning to support more diverse population-based studies, it is also essential for local and national governments to cooperatively fund and support the creation of regional biobanks. These biobanks can leverage the existing health care infrastructure to enroll and phenotype participants, the most difficult, costly, and time-consuming step in genetics studies. Sequencing and genotyping, by comparison, costs as little as tens of dollars per participant, and large sequencing centers are capable of genotyping thousands of individuals per day. While the creation of large biobanks is an expensive and daunting task, the UK Biobank has shown that they facilitate research at scales never before seen.

Finally, further study of individuals from Chocó is necessary to understand the genetic architecture of disease fully. A proposed ChocoGEN2 would encompass a larger prospective cohort along with deep phenotyping to provide. This would enable association studies and other analyses that rely on individual-level phenotype data. This thesis has shown that rudimentary population precision health recommendations can be made with

the existing data. Expanding the depth of data available to researchers would enable even more public health benefits and decrease health disparities.

APPENDIX A. SUPPLEMENTARY TABLES FOR CHAPTER 3

Table 5. SNP information for SNPs in Figure 2

Effect Population	SNP	REF	ALT	Ref Freq ANT	Ref Freq CHO	Alt Freq ANT	Alt Freq CHO	Effect Allele	Effect Complement	OR	F_{ST}	P-value
Antioquia	rs16891982	C	G	0.361702	0.93617	0.638298	0.06383	C	G	0.31	0.362	3.92E-10
	rs11894081	G	T	0.202128	0.797872	0.797872	0.202128	T	A	1.22	0.355	8.42E-10
	rs1137	T	C	0.260646	0.792553	0.739354	0.207447	C	G	0.12	0.284	8.90E-07
	rs4474514	G	A	0.164894	0.691489	0.835106	0.308511	A	T	3.07	0.283	9.72E-07
	rs2472649	A	G	0.244681	0.765957	0.755319	0.234043	G	C	1.095	0.272	2.50E-06
	rs995030	A	G	0.159574	0.670213	0.840426	0.329787	G	C	2.69	0.269	3.22E-06
	rs1426654	A	G	0.718085	0.207447	0.281915	0.792553	A	T	0.484	0.262	5.75E-06
	rs12940030	T	C	0.680851	0.175532	0.319149	0.824468	T	A	0.08	0.261	6.23E-06
	rs1667394	C	T	0.425532	0.882979	0.574468	0.117021	A	T	4.94	0.231	6.30E-05
	rs224333	G	A	0.680851	0.212766	0.319149	0.787234	G	C	0.0363	0.222	1.20E-04
Chocó	rs28777	C	A	0.345745	0.856383	0.654255	0.143617	C	G	0.46	0.272	3.00E-06
	rs1298637	C	T	0.718085	0.191489	0.281915	0.808511	A	T	5.972	0.280	1.52E-06
	rs9623117	T	C	0.851064	0.324468	0.148936	0.675532	C	G	1.18	0.286	9.03E-07
	rs1834640	A	G	0.781915	0.239362	0.218085	0.760638	G	C	12.5	0.294	4.43E-07
	rs509360	A	G	0.239362	0.797872	0.760638	0.202128	A	T	0.0031	0.312	8.35E-08
	rs10007810	G	A	0.765957	0.202128	0.234043	0.797872	A	T	1.2	0.318	4.69E-08
	rs11814448	A	C	0.904255	0.345745	0.095745	0.654255	C	G	1.26	0.333	1.06E-08
	rs12074934	T	G	0.941489	0.356383	0.058511	0.643617	C	G	15.23	0.376	1.04E-10
	rs2814778	T	C	0.925532	0.18617	0.074468	0.81383	C	G	1.35	0.554	1.74E-21

Table 5 continued

Effect Population	SNP	Trait	Mapped gene(s)	Annotation
Antioquia	rs16891982	Skin sensitivity to sun	SLC24A5	C/G protective against malignant melanoma
	rs11894081	Crohn's disease		Unspecified rate increase
	rs1137	Myopia pathological	SEMA4F	Axon growth disorders
	rs4474514	Testicular cancer	KITLG	3x rate increase per A
	rs2472649	Inflammatory bowel disease	Intronic	Unknown function
	rs995030	Testicular germ cell tumor	KITLG	2.5x rate increase per G
	rs1426654	Body mass index	SLC24A5	Lighter skin color per A allele
	rs12940030	Corneal structure	HS3ST3B1-PMP22	Corneal thickness
	rs1667394	Blond vs. brown hair color		A allele 4.4x more likely to be blond
	rs224333	Waist-to-hip ratio adjusted for body mass index		G increases adipose tissue growth
Chocó	rs28777	Black(vs. blond) hair color	SLC24A5	Darker hair for each C allele
	rs1298637	Nicotine use	GLIS1	Increased nicotine use
	rs9623117	Prostate cancer	TNRC6B	Increased prostate cancer risk in African Americans
	rs1834640	Skin pigmentation - dark	SLC24A5	Darker skin pigmentation for each G
	rs509360	Trans fatty acid levels	FEN1	Increased ALA and LA levels from PUFA pathway
	rs10007810	Longevity 90 years and older	LIMCH1	Small increase in longevity
	rs11814448	Breast cancer	ADIPOR1P1, DNAJC1	Unknown function
	rs12074934	Diisocyanate-induced asthma	OR10J3	Unknown function
	rs2814778	Resistance to *Plasmodium vivax*	DARC (ACKR1)	Duffy blood group is protective against malaria

Table 6. Concordance between predicted and observed trait differences between Antioquia and Chocó.

Concordant prediction = 21, Discordant prediction = 4, p-value = 0.000910521

Trait	Affected Population	Sign
Allergic sensitization	Antioquia	+
Blond (vs. brown) hair color	Antioquia	+
BMI change over time	Antioquia	-
Body mass index	Antioquia	-
Corneal structure	Antioquia	+
Crohn's disease	Antioquia	+
Height	Antioquia	+
Inflammatory bowel disease	Antioquia	+
Myopia (pathological)	Antioquia	+
Skin sensitivity to sun	Antioquia	+
Stroke, ischemic	Antioquia	+
Testicular cancer	Antioquia	+
Testicular germ cell tumar	Antioquia	+
WHR-adjusted BMI	Antioquia	-
Alzheimer's disease, late-onset	Chocó	U
Black (vs. blond) hair color	Chocó	+
Breast cancer	Chocó	+
Diisocyanate-induced asthma	Chocó	+
Eye color (Brown vs. Blue/Green)	Chocó	+
Longevity, 90 years and older	Chocó	U
Mortality in heart failure	Chocó	+
Nicotine use	Chocó	-
Prostate cancer	Chocó	+
Resistance to Plasmodium vivax	Chocó	+
Skin saturation	Chocó	+
Skin pigmentation - dark	Chocó	+
Trans fatty acid levels	Chocó	+

Table 7. GWAS Catalog trait PRS differences between Chocó and Antioquia

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
β2-Glycoprotein I β2-GPI plasma levels	-7.074468	5.34E-03	Chocó	European
Abdominal aortic aneurysm	-9.425211	3.39E-11	Chocó	NR, European
Activated partial thromboplastin time	12.471505	2.94E-11	Antioquia	East Asian
Advanced glycation end-product levels	-7.180851	1.77E-02	Chocó	NR
Adverse response to chemotherapy in breast cancer alopecia cyclophosphamide+doxorubicin+-over--5FU	10.460993	3.44E-07	Antioquia	East Asian
Adverse response to chemotherapy in breast cancer alopecia docetaxel	-15.2039	2.61E-09	Chocó	East Asian
Adverse response to chemotherapy in breast cancer alopecia paclitaxel	-10.85993	2.65E-10	Chocó	East Asian
Adverse response to chemotherapy neutropenia-over-leucopenia all anthracycline-based drugs	-10.66489	4.71E-05	Chocó	East Asian
Adverse response to chemotherapy neutropenia-over-leucopenia all platinum-based drugs	-11.0195	1.32E-11	Chocó	East Asian
Adverse response to chemotherapy neutropenia-over-leucopenia gemcitabine	5.9574468	3.81E-02	Antioquia	East Asian
Age at smoking initiation in chronic obstructive pulmonary disease	11.931889	1.30E-11	Antioquia	European
Aggressive periodontitis	-11.94149	3.04E-05	Chocó	European, Other
Aging time to event	8.2117528	4.96E-04	Antioquia	European
AIDS	-10.94225	5.34E-03	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
AIDS	-10.94225	8.42E-07	Chocó	European
Airway imaging phenotypes	-13.89628	9.62E-17	Chocó	African American or Afro-Caribbean, European
Airway responsiveness in chronic obstructive pulmonary disease	-8.079217	3.70E-10	Chocó	European
Alanine aminotransferase ALT levels after remission induction therapy in actute lymphoblastic leukemia ALL	-3.790717	4.68E-07	Chocó	African American or Afro-Caribbean, European, Hispanic or Latin American, Other
Alcohol consumption	8.8219693	5.54E-08	Antioquia	European
Alcohol consumption drinks per week	16.755319	3.83E-03	Antioquia	East Asian, European, Hispanic or Latin American, African American or Afro-Caribbean
Alcohol consumption heavy vs. light-over-non-drinkers	-31.91489	3.62E-06	Chocó	European
Alcohol consumption in current drinkers	6.8009119	1.06E-03	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Alcohol dependence	4.7969609	8.50E-04	Antioquia	African American or Afro-Caribbean, European
Alcohol dependence or chronic alcoholic pancreatitis or alcohol-related liver cirrhosis	10.992908	4.46E-03	Antioquia	European
Alcohol dependence symptom count	-5.717778	2.80E-02	Chocó	African American or Afro-Caribbean, European
Alcohol use disorder total score	9.8210832	3.00E-09	Antioquia	European
Alcoholism alcohol use disorder factor score	15.780142	4.00E-13	Antioquia	European
Allergic disease asthma, hay fever or eczema	2.9188052	6.85E-10	Antioquia	European
Alzheimer's disease	5.9940768	8.12E-04	Antioquia	NR, African unspecified, European
Alzheimer's disease biomarkers	-8.018521	2.80E-04	Chocó	European
Amyotrophic lateral sclerosis	-5.123895	1.92E-08	Chocó	European
Angiotensin-converting enzyme activity	11.968085	3.14E-03	Antioquia	East Asian
Ankle-brachial index	-26.32979	1.28E-08	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Anterior chamber depth	22.87234	7.23E-03	Antioquia	East Asian, South Asian, South East Asian
Anterior cruciate ligament rupture	7.535461	6.08E-04	Antioquia	East Asian, South East Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Anti-saccade response	-3.786784	1.79E-03	Chocó	European
Antitragus size	20.212766	3.22E-03	Antioquia	Hispanic or Latin American
Anxiety and major depressive disorder	-27.12766	1.21E-08	Chocó	European
Anxiety disorder	-5.452665	5.07E-04	Chocó	African American or Afro-Caribbean, European
Area under the curve of insulin levels	10.638298	1.86E-03	Antioquia	European
Arthritis juvenile idiopathic	8.9361702	3.01E-07	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Asparaginase hypersensitivity in acute lymphoblastic leukemia	-8.999493	4.20E-04	Chocó	European, Hispanic or Latin American, African unspecified, Asian unspecified, Other
Asthma bronchodilator response	-18.35106	7.37E-09	Chocó	European, Hispanic or Latin American
Asthma childhood onset	-6.391627	1.59E-04	Chocó	East Asian, European, Hispanic or Latin American, African unspecified, African American or Afro-Caribbean
Asthma moderate or severe	8.4637312	2.79E-15	Antioquia	European
Atopic dermatitis	-3.337761	7.52E-03	Chocó	East Asian
Atopic march	-9.840426	1.15E-05	Chocó	European
Attention deficit hyperactivity disorder	-3.119673	1.63E-02	Chocó	European
Attention deficit hyperactivity disorder or cannabis use	-6.833804	4.18E-03	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Autism spectrum disorder	-15.86879	3.98E-09	Chocó	European, Other, NR
Autism spectrum disorder or schizophrenia	5.6996313	1.36E-14	Antioquia	European, Other, NR
Autoimmune thyroid diseases Graves disease or Hashimoto's thyroiditis	8.7765957	1.30E-04	Antioquia	NR, European
Basal cell carcinoma	8.8956712	1.69E-15	Antioquia	European
Basal metabolic rate	23.93617	1.64E-03	Antioquia	East Asian
B-cell malignancies chronic lymphocytic leukemia, Hodgkin lymphoma or multiple myeloma pleiotropy	-12.43517	1.91E-15	Chocó	European
Behcet's disease	-12.14539	8.95E-11	Chocó	East Asian
Biliary atresia	-13.67908	4.77E-08	Chocó	European
Bilirubin levels	-4.516016	1.48E-02	Chocó	East Asian
Biochemical measures	-5.472276	1.81E-05	Chocó	European
Bipolar I disorder	-18.35106	2.53E-10	Chocó	East Asian
Black vs. blond hair color	-19.7695	1.49E-26	Chocó	European
Black vs. red hair color	-20	3.20E-25	Chocó	European
Blond vs. brown hair color	9.3085106	1.32E-02	Antioquia	European
Blond vs. brown-over-black hair color	-2.681003	3.68E-05	Chocó	European
Blood and toenail selenium levels	-12.2824	2.10E-15	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Blood glucose levels	11.170213	7.55E-03	Antioquia	European
Blood osmolality transformed sodium	-5.522538	1.59E-10	Chocó	South Asian, European, African American or Afro-Caribbean
Blood pressure measurement low sodium intervention	-11.43617	6.21E-03	Chocó	East Asian
Blood pressure traits multi-trait analysis	-9.931611	3.37E-04	Chocó	East Asian, African unspecified, Hispanic or Latin American, European
Blue vs. green eyes	10.815603	6.35E-04	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
BMI in smokers	-14.45774	8.00E-13	Chocó	East Asian, South Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
BMI smoking interaction	-39.89362	3.88E-24	Chocó	East Asian, South Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Body fat mass	-27.12766	2.29E-05	Chocó	European
Body mass index smoking years interaction	-8.333333	4.83E-03	Chocó	African American or Afro-Caribbean, Hispanic or Latin American

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Body mass index age interaction	-21.80851	2.29E-04	Chocó	NR, European
Body mass index in asthmatics	-9.388298	1.62E-03	Chocó	NR, European, Hispanic or Latin American, Other
Body mass index in non-asthmatics	-9.397163	1.87E-07	Chocó	NR, European, Hispanic or Latin American, Other
Body mass index in physically active individuals	6.7110897	1.14E-08	Antioquia	East Asian, European, South Asian, African American or Afro-Caribbean
Body mass index SNP x SNP interaction	7.7001013	6.71E-03	Antioquia	East Asian
Body mass index x sex x age interaction 4df test	-7.644262	1.45E-04	Chocó	European
Bone fracture in osteoporosis	-7.679838	6.63E-05	Chocó	African American or Afro-Caribbean
Bone mineral density femoral neck in inflammatory bowel disease	-12.6773	3.72E-09	Chocó	East Asian
Bone mineral density hip	-6.54133	4.51E-10	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Bone mineral density paediatric, skull	-30.31915	3.42E-12	Chocó	European, Greater Middle Eastern (Middle Eastern, North African or Persian), Other
Bone ultrasound measurement broadband ultrasound attenuation	-6.530286	2.96E-05	Chocó	European
Borderline personality disorder	-4.232103	2.57E-02	Chocó	European
Brain structure	-21.01064	2.14E-06	Chocó	European
Breast cancer	2.7219777	5.45E-23	Antioquia	East Asian, European
Breast Cancer in BRCA1 mutation carriers	-5.912749	1.52E-04	Chocó	European
Bronchopulmonary dysplasia in preterm infants	-22.25177	2.75E-05	Chocó	African American or Afro-Caribbean, European, Hispanic or Latin American
Brown vs. black hair color	-7.196609	3.71E-06	Chocó	European
Brugada syndrome	22.695035	5.51E-15	Antioquia	East Asian, European
B-type natriuretic peptide to N-terminal pro B-type natriuretic peptide ratio	11.258865	1.44E-03	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Caffeine metabolism plasma 1,3-dimethylxanthine theophylline level	7.712766	2.15E-04	Antioquia	European
Caffeine metabolism plasma 1,7-dimethylxanthine paraxanthine to 1,3,7-trimethylxanthine caffeine ratio	7.1032141	4.76E-12	Antioquia	European
Calcium levels	4.5152896	7.26E-05	Antioquia	East Asian
Cannabis use	8.5387116	1.05E-06	Antioquia	European
Cannabis use initiation	-4.872758	1.03E-03	Chocó	European
Cardiac repolarization	12.012411	1.72E-06	Antioquia	European
Cardiac Troponin-T levels	-6.433842	2.20E-05	Chocó	African American or Afro-Caribbean, European
Cardiovascular disease risk factors	8.6992263	1.06E-06	Antioquia	European
Carotid intima media thickness	-7.765957	1.74E-02	Chocó	East Asian
Cataracts in type 2 diabetes	-18.61702	1.18E-04	Chocó	East Asian
Caudate activity during reward	5.5849459	1.90E-09	Antioquia	NR
Caudate nucleus volume	14.095745	5.41E-04	Antioquia	European
Celiac disease	2.3406319	6.02E-03	Antioquia	South Asian, European
Celiac disease or Rheumatoid arthritis	9.6757852	5.57E-05	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Central corneal thickness	2.6807793	1.03E-02	Antioquia	Asian unspecified, European
Cerebral amyloid angiopathy	-10.94858	1.21E-03	Chocó	NR
Cerebrospinal AB1-42 levels in Alzheimer's disease dementia	-13.56383	1.31E-03	Chocó	European
Cerebrospinal AB1-42 levels in normal cognition	-5.162261	6.89E-04	Chocó	European
Cerebrospinal fluid α-synuclein levels	-10.07979	1.55E-07	Chocó	European
Cerebrospinal T-tau levels	-4.251046	2.93E-06	Chocó	NR
Cerivastatin-induced rhabdomyolysis	-22.87234	4.72E-04	Chocó	NR, European
Change in intraocular pressure in response to steroid treatment triamcinolone acetonide	-21.2766	4.52E-02	Chocó	African American or Afro-Caribbean, European
Childhood onset systemic lupus erythematosus	17.021277	3.60E-02	Antioquia	East Asian
Cholesterol and Triglycerides	-25	3.42E-03	Chocó	European
Cholesterol, total	-3.078207	1.51E-07	Chocó	Asian unspecified, Sub-Saharan African, European, NR
Chronic hepatitis B infection	7.7611219	1.23E-09	Antioquia	East Asian
Chronic kidney disease	-6.084721	7.92E-07	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Chronic kidney disease and serum creatinine levels	14.539007	1.40E-06	Antioquia	European
Chronic lymphocytic leukemia	-3.034269	1.90E-02	Chocó	European, NR, European
Chronic myeloid leukemia	-21.80851	3.08E-05	Chocó	East Asian, European
Chronotype	-7.163005	1.27E-10	Chocó	European
Circulating fibroblast growth factor 23 levels	5.0075368	1.77E-02	Antioquia	African unspecified, European
circulating leptin levels	-18.08511	1.69E-03	Chocó	European
Circulating myeloperoxidase levels serum	-19.77837	3.99E-13	Chocó	European
Circulating phylloquinone levels	10.01773	2.23E-02	Antioquia	European
Circulating vasoactive peptide levels	-13.03191	3.44E-02	Chocó	European
Classic bladder exstrophy	-5.8274	5.69E-07	Chocó	European
Cleft lip	-27.12766	1.81E-04	Chocó	Asian unspecified, European, South Asian, Other

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Cleft palate	-25.26596	2.49E-13	Chocó	Hispanic or Latin American, African unspecified, Asian unspecified, European, NR
Clozapine-induced agranulocytosis	5.5163396	2.50E-14	Antioquia	European
Coffee consumption	-8.766253	1.74E-07	Chocó	East Asian
Cognitive decline rate in late mild cognitive impairment	2.6944457	1.08E-02	Antioquia	NR, European
Cognitive performance	1.6577381	2.95E-03	Antioquia	European
Cognitive performance MTAG	1.6805255	3.73E-14	Antioquia	European
Common carotid intima-media thickness	8.4530142	4.34E-06	Antioquia	African American or Afro-Caribbean
Common carotid intima-media thickness in HIV infection	15.21361	7.92E-16	Antioquia	African American or Afro-Caribbean
Complement C3 and C4 levels	-6.170213	2.04E-03	Chocó	East Asian
Conduct disorder symptom count	-10.72802	1.57E-09	Chocó	African American or Afro-Caribbean, European, Other

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Congenital left-sided heart lesions	-16.2234	9.57E-11	Chocó	European
Corneal astigmatism	-4.965165	2.43E-07	Chocó	European
Coronary artery disease in diabetes	-17.02128	8.35E-07	Chocó	European
Coronary artery disease in type 1 diabetes	4.8260531	1.93E-04	Antioquia	NR, European
Coronary restenosis	-22.34043	2.97E-03	Chocó	European
Cortical thickness	17.553191	3.14E-03	Antioquia	European
Cotinine glucuronidation	6.4705758	1.52E-04	Antioquia	East Asian, Oceanian, African American or Afro-Caribbean, European, Hispanic or Latin American
Craniofacial microsomia	10.91218	9.03E-10	Antioquia	East Asian
C-reactive protein levels	-5.489546	1.31E-06	Chocó	East Asian
C-reactive protein levels or HDL-cholesterol levels pleiotropy	5.8795593	4.34E-05	Antioquia	NR
Creatinine levels	4.1578014	3.09E-03	Antioquia	East Asian
Crohn's disease and celiac disease	-25.79787	1.89E-09	Chocó	European
Crohn's disease and psoriasis	-17.02128	1.10E-04	Chocó	European
Crohn's disease-related phenotypes	-7.765957	2.36E-02	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Current cigarettes per day in chronic obstructive pulmonary disease	8.1781915	3.17E-06	Antioquia	African American or Afro-Caribbean, European, NR
Cutaneous malignant melanoma	-9.362911	2.64E-06	Chocó	NR, European
Cutaneous squamous cell carcinoma	19.379433	5.79E-18	Antioquia	European
Cystic fibrosis severity	16.489362	1.04E-05	Antioquia	NR, European
Cystic fibrosis-related diabetes	-29.78723	1.10E-05	Chocó	NR
Dehydroepiandrosterone sulphate levels	6.7354779	1.07E-03	Antioquia	NR, European
Delta-5 desaturase activity response to n3-polyunsaturated fat supplement	-10.60284	1.44E-13	Chocó	NR
Delta-6 desaturase activity	-9.521277	2.58E-04	Chocó	East Asian
Dementia and core Alzheimer's disease neuropathologic changes	-5.489957	1.62E-03	Chocó	NR
Dentate gyrus granule cell layer volume	-9.83156	5.43E-05	Chocó	European
Dentate gyrus molecular layer volume corrected for total hippocampal volume	-27.65957	1.20E-05	Chocó	European
Depression	2.722528	8.72E-07	Antioquia	European
Developmental dysplasia of the hip	13.696809	1.17E-03	Antioquia	European
Diabetes gestational	-30.85106	6.12E-13	Chocó	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Diabetic kidney disease	-3.917733	1.08E-03	Chocó	African American or Afro-Caribbean, European, Hispanic or Latin American, Native American
Diastolic blood pressure x alcohol consumption light vs heavy interaction 2df test	3.9842399	6.02E-03	Antioquia	African American or Afro-Caribbean, Asian unspecified, European, African unspecified, Hispanic or Latin American
Digit length ratio	10.967579	6.23E-08	Antioquia	European
Digit length ratio right hand	6.5634498	1.17E-02	Antioquia	NR, European
Diisocyanate-induced asthma	-3.408175	1.07E-31	Chocó	European
Disease progression in age-related macular degeneration	-3.863861	1.20E-06	Chocó	European
Disease-free survival in breast cancer	-13.38652	1.03E-03	Chocó	East Asian
Disrupted circadian rhythm low relative amplitude of rest-activity cycles	15.780142	1.90E-07	Antioquia	European
Diverticular disease	3.2349325	1.40E-10	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
DNA methylation parent-of-origin	8.6879433	5.28E-03	Antioquia	European
Ear morphology	13.994428	4.60E-14	Antioquia	Hispanic or Latin American
Economic and political preferences	-8.87633	1.78E-11	Chocó	European
Economic and political preferences environmentalism	5.7537309	9.77E-04	Antioquia	European
Educational attainment MTAG	1.8611321	2.09E-25	Antioquia	European
Educational attainment years of education	2.1101014	3.09E-34	Antioquia	European
Electrocardiographic conduction measures	12.047239	9.76E-09	Antioquia	European
Elevated fasting plasma glucose	-4.361702	2.05E-03	Chocó	East Asian
Empathy quotient	8.3510638	1.63E-02	Antioquia	European
Emphysema distribution in smoking	-9.086879	1.01E-06	Chocó	African American or Afro-Caribbean, European, NR
Emphysema-related traits	7.9787234	1.15E-03	Antioquia	European
Endometrial cancer endometrioid histology	5.8271293	4.05E-08	Antioquia	European
End-stage renal disease	-25	1.04E-02	Chocó	African American or Afro-Caribbean
Eosinophil counts	3.1882989	1.19E-12	Antioquia	East Asian
Eosinophil percentage of granulocytes	1.8581133	4.55E-03	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Eosinophil percentage of white cells	2.3195984	3.48E-07	Antioquia	European
Epilepsy generalized	-8.546099	2.05E-07	Chocó	European
Epirubicin-induced leukopenia	16.489362	4.97E-07	Antioquia	East Asian
Epithelial ovarian cancer	4.9057505	6.66E-09	Antioquia	European
Erythrocyte sedimentation rate	-35.6383	6.15E-09	Chocó	European
Esophageal adenocarcinoma x smoking interaction	-9.840426	1.05E-02	Chocó	European
Esophageal cancer	6.2886272	4.99E-04	Antioquia	East Asian
Esophageal cancer squamous cell	10.514184	1.97E-05	Antioquia	East Asian
Estradiol levels	-9.299645	1.86E-05	Chocó	European
Estrone-over-androstenedione ratio in resected early stage-receptor positive breast cancer	6.0949925	3.41E-03	Antioquia	Asian unspecified, European, African American or Afro-Caribbean
Eudaimonic well-being	5.5834847	6.34E-04	Antioquia	European
Event-related brain oscillations	8.5992908	3.19E-05	Antioquia	African American or Afro-Caribbean, European, Other
Extraversion	15.780142	2.52E-05	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Eye color	-11.36229	3.95E-13	Chocó	European
Facial morphology factor 12, vertical position of sublabial sulcus relative to central midface	8.0350811	2.66E-03	Antioquia	European
Facial morphology factor 17, height of vermillion upper lip	5.4252633	3.57E-03	Antioquia	European
Facial morphology factor 18	-5.044264	8.29E-03	Chocó	European
Facial morphology factor 23	6.8234971	3.14E-03	Antioquia	European
Facial morphology factor 3, length of philtrum	-9.267139	1.89E-11	Chocó	European
Facial morphology factor 5, width of mouth relative to central midface	-7.123227	7.54E-03	Chocó	European
Fasting blood glucose	-6.391542	2.43E-07	Chocó	African unspecified, European
Fasting blood glucose BMI interaction	-3.982747	2.20E-03	Chocó	European
Fear of minor pain	-17.02145	1.46E-36	Chocó	Asian unspecified, Native American, Other, African American or Afro-Caribbean, European, NR, Hispanic or Latin American

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Fear of severe pain	6.7933131	3.69E-03	Antioquia	Asian unspecified, Native American, Other, African American or Afro-Caribbean, European, NR, Hispanic or Latin American
Feeling lonely	-10.21277	4.39E-07	Chocó	European
Feeling nervous	3.4290164	8.61E-07	Antioquia	European
Feeling worry	3.9579882	2.78E-07	Antioquia	European
Fibrinogen levels smoking status, alcohol consumption or body mass index interaction	-18.88298	5.19E-05	Chocó	European
Folding of antihelix	28.191489	6.42E-13	Antioquia	Hispanic or Latin American
Formal thought disorder in schizophrenia	#VALUE!	1.52E-02	#VALUE!	European
Fractional excretion of metabolites in chronic kidney disease	12.5	1.77E-04	Antioquia	European
Fractional exhaled nitric oxide childhood	-9.046353	9.24E-06	Chocó	European, Hispanic or Latin American, Other

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Fractures	5.0040814	8.46E-03	Antioquia	European
Freckling	-8.156028	2.13E-03	Chocó	European
Frontotemporal dementia	7.8492371	1.43E-04	Antioquia	European
Gallstone disease	-14.30851	2.18E-09	Chocó	African American or Afro-Caribbean, European, Hispanic or Latin American
Gamma glutamyl transferase levels	7.9787234	4.27E-03	Antioquia	East Asian
Gene methylation in lung tissue	-30.31915	5.58E-14	Chocó	European
General cognitive ability	1.1833135	4.77E-04	Antioquia	European
Gestational age at birth in labor-initiated deliveries child effect	4.061128	3.59E-03	Antioquia	European
Gestational age at birth in premature rupture of membrane-initiated deliveries child effect	-19.23759	1.33E-18	Chocó	European
Glaucoma high intraocular pressure	-12.18085	4.60E-10	Chocó	East Asian, European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Glaucoma primary open-angle	-6.276442	5.06E-13	Chocó	South Asian, East Asian, Other admixed ancestry, African American or Afro-Caribbean, European, African unspecified, Hispanic or Latin American
Glomerular filtration rate creatinine	4.4173314	3.68E-17	Antioquia	European
Glomerular filtration rate in non diabetics creatinine	5.4387657	9.74E-07	Antioquia	Asian unspecified, African unspecified, European
Glycemic traits pregnancy	23.847518	6.72E-18	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American, South East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Granulocyte percentage of myeloid white cells	2.9208545	1.10E-12	Antioquia	European
Graves' disease	5.6138389	6.05E-07	Antioquia	European
Gut microbiota bacterial taxa	6.5950614	2.72E-09	Antioquia	European
Hair color	5.8992957	1.02E-04	Antioquia	Hispanic or Latin American
Hair shape	8.9228723	5.09E-10	Antioquia	East Asian, European, Other, South Asian, Greater Middle Eastern (Middle Eastern, North African or Persian), Hispanic or Latin American
HDL Cholesterol - Triglycerides HDLC-TG	-14.71631	2.58E-11	Chocó	European
HDL cholesterol levels	5.2301837	1.51E-10	Antioquia	NR, European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Heart rate response to recovery post exercise 10 sec	-5.700727	1.66E-04	Chocó	Other admixed ancestry, Other, NR, Asian unspecified, European, African unspecified
Heart rate response to recovery post exercise 20 sec	-6.159204	4.00E-06	Chocó	Other admixed ancestry, Other, NR, Asian unspecified, European, African unspecified
Heart rate response to recovery post exercise 30 sec	-6.042876	3.43E-04	Chocó	Other admixed ancestry, Other, NR, Asian unspecified, European, African unspecified

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Heart rate response to recovery post exercise 40 sec	-5.456157	2.00E-03	Chocó	Other admixed ancestry, Other, NR, Asian unspecified, European, African unspecified
Heart rate variability traits pvRSA-over-HF	19.414894	4.57E-14	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American
Heart rate variability traits RMSSD	9.047619	4.24E-04	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American
Hedonic well-being	11.968085	4.58E-05	Antioquia	European
Heel bone mineral density	-2.20024	1.27E-30	Chocó	European
Hematocrit	3.2300364	4.08E-12	Antioquia	East Asian
Hematology traits	8.9866753	5.12E-12	Antioquia	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Hemoglobin	-7.609338	4.12E-07	Chocó	East Asian
Hemoglobin A1c levels	-23.04965	7.65E-09	Chocó	African unspecified, European
Hemoglobin A2 levels in sickle cell anemia	-10.90426	1.53E-04	Chocó	African American or Afro-Caribbean, East Asian
Hemoglobin levels	-7.87234	6.81E-03	Chocó	NR
Hen's egg allergy	14.095745	5.73E-04	Antioquia	European
Hepatocellular carcinoma in hepatitis B infection	-12.76596	1.90E-02	Chocó	East Asian, South East Asian
Hepatocyte growth factor levels	-7.052841	5.29E-08	Chocó	European
Hepcidin-over-transferrin saturation ratio	14.911348	2.96E-10	Antioquia	NR, European
Heschl's gyrus morphology	3.290158	1.60E-02	Antioquia	NR, European
High altitude adaptation	6.3012918	9.12E-03	Antioquia	East Asian
Highest math class taken	-2.514678	5.78E-14	Chocó	European
Highest math class taken MTAG	1.8900323	5.33E-26	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Hip circumference	17.698835	1.37E-17	Antioquia	South Asian, African American or Afro-Caribbean, Asian unspecified, European, African unspecified
Hippocampal subfield CA3 volume	-8.156028	4.27E-02	Chocó	European
Hippocampal subfield CA4 volume	-9.83156	5.43E-05	Chocó	European
Hippocampal tail volume corrected for total hippocampal volume	-10.99291	7.98E-06	Chocó	European
Hippocampal volume in Alzheimer's disease dementia	8.643617	7.92E-06	Antioquia	European
Hirschsprung disease	11.347518	5.08E-17	Antioquia	European
HIV-associated dementia	6.7375887	3.62E-02	Antioquia	European
Hoarding	10.815603	2.34E-03	Antioquia	European
Homoarginine levels	25.531915	5.26E-04	Antioquia	European
Homocysteine levels	-7.833656	8.35E-08	Chocó	NR

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Hypertension	-2.679299	3.40E-02	Chocó	South Asian, European, Hispanic or Latin American, African American or Afro-Caribbean
Hypertriglyceridemia	15.970111	1.21E-11	Antioquia	Greater Middle Eastern (Middle Eastern, North African or Persian)
Idiopathic inflammatory myopathy	15.691489	8.77E-03	Antioquia	European
Idiopathic membranous nephropathy	-2.852744	2.49E-03	Chocó	European
IgE levels	8.2692618	1.28E-08	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American
IgE levels in asthmatics D.f. specific	-23.40426	1.72E-02	Chocó	East Asian
IgG digalactosylation phenotypes multivariate analysis	-11.8617	6.30E-06	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
IgG fucosylation phenotypes multivariate analysis	-7.819149	2.70E-02	Chocó	European
IgG glycosylation patterns	10.34824	7.64E-04	Antioquia	European
IgG N-glycosylation phenotypes multivariate analysis	-8.35233	4.77E-04	Chocó	European
IgG sialylation phenotypes multivariate analysis	-10.22036	6.99E-05	Chocó	European
Immune reponse to smallpox secreted IFN-alpha	1.5268685	4.43E-02	Antioquia	African American or Afro-Caribbean, European
Immune reponse to smallpox secreted IL-12p40	-4.853723	3.16E-05	Chocó	African American or Afro-Caribbean, European
Immune reponse to smallpox secreted IL-1beta	-3.239845	6.25E-07	Chocó	African American or Afro-Caribbean, European
Immune reponse to smallpox secreted IL-2	-2.087133	2.93E-03	Chocó	African American or Afro-Caribbean, European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Incident atrial fibrillation	14.716312	4.49E-06	Antioquia	East Asian, European, NR, African American or Afro-Caribbean, European, Hispanic or Latin American
Incident coronary heart disease	-20.74468	3.01E-08	Chocó	African American or Afro-Caribbean, European
Incident myocardial infarction	-13.20922	1.71E-04	Chocó	African American or Afro-Caribbean, European
Inflammatory biomarkers	-7.408815	3.10E-02	Chocó	European
Inflammatory biomarkers in Kawasaki disease	-26.06383	1.27E-06	Chocó	East Asian
Inguinal hernia	-17.28723	4.70E-06	Chocó	European
Inhibitory control	7.0478723	8.80E-05	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Insulin secretion rate	-32.44681	2.00E-06	Chocó	European, Hispanic or Latin American, Native American
Insulin-like growth factors	10.54078	1.14E-02	Antioquia	European
Intelligence MTAG	0.9187886	3.27E-02	Antioquia	NR, European
Interleukin-13 levels	4.4854968	1.30E-02	Antioquia	European
Interleukin-17 levels	-4.316933	7.32E-04	Chocó	European
Interleukin-1-beta levels	-6.83574	3.67E-04	Chocó	European
Interleukin-4 levels	-8.449532	1.76E-07	Chocó	European
Interleukin-5 levels	-8.739657	9.27E-08	Chocó	European
Interleukin-6 levels	3.7557128	2.76E-02	Antioquia	African American or Afro-Caribbean, Sub-Saharan African
Interstitial lung disease	-4.979662	2.87E-03	Chocó	European
Intraocular pressure	-1.992853	1.11E-07	Chocó	European
Iron levels	-22.34043	1.32E-02	Chocó	NR, European
Iron status biomarkers ferritin levels	-11.35702	2.66E-12	Chocó	Hispanic or Latin American

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Irritable mood	3.1622038	7.80E-04	Antioquia	European
Ischemic stroke small artery occlusion	44.148936	1.99E-15	Antioquia	East Asian, South Asian, African American or Afro-Caribbean, African unspecified, European, Hispanic or Latin American
Ischemic stroke small-vessel	-10.6383	1.70E-02	Chocó	East Asian, South Asian, African American or Afro-Caribbean, Asian unspecified, European, Hispanic or Latin American
Isovolumetric relaxation time	14.804965	1.47E-07	Antioquia	European
Kidney disease end stage renal disease vs non-end stage renal disease in type 1 diabetes	-26.59574	5.16E-04	Chocó	NR, European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
L-arginine levels	-12.5	1.48E-02	Chocó	South Asian, European, NR
Late-onset myasthenia gravis	8.0870061	1.73E-03	Antioquia	European
LDL cholesterol levels	-3.307453	2.00E-04	Chocó	NR, European
Leisure-time exercise behaviour	-10.50532	2.41E-03	Chocó	East Asian
Leprosy	3.4959782	8.06E-03	Antioquia	East Asian
Lipoprotein a levels	-3.125087	6.30E-03	Chocó	European
Lipoprotein phospholipase A2 activity in cardiovascular disease	-4.604969	7.23E-03	Chocó	European, Asian unspecified, African American or Afro-Caribbean, Greater Middle Eastern (Middle Eastern, North African or Persian), Oceanian, Native American, Other admixed ancestry,
Lipoproteina levels adjusted for apolipoproteina isoforms	4.1113319	7.37E-06	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Lipoprotein-associated phospholipase A2 activity and mass	-7.162315	7.74E-10	Chocó	European
Lipoprotein-associated phospholipase A2 activity change in response to statin therapy	8.0851064	9.76E-04	Antioquia	European
Liver injury in anti-tuberculosis drug treatment	22.606383	4.10E-08	Antioquia	Sub-Saharan African
Lobe size	10.505319	7.52E-04	Antioquia	Hispanic or Latin American
Loneliness	-7.625457	6.62E-05	Chocó	European
Loneliness MTAG	-8.471582	2.25E-06	Chocó	European
Longevity 90 years and older	-15.02026	1.53E-13	Chocó	European
Low tan response	-15.11195	1.01E-43	Chocó	NR, European
Low white blood cell count	-73.93617	9.44E-45	Chocó	African American or Afro-Caribbean
Lower body strength	12.327761	3.41E-08	Antioquia	NR, European
Lung adenocarcinoma	4.2362955	3.92E-08	Antioquia	European
Lung cancer in ever smokers	3.6223337	1.35E-09	Antioquia	European
Lung cancer smoking interaction	-10.6383	1.76E-02	Chocó	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Lung function FEV1	5.3203936	2.68E-07	Antioquia	East Asian, Hispanic or Latin American, African unspecified, European
Lung function FEV1-over-FVC	3.627147	3.66E-07	Antioquia	East Asian, Hispanic or Latin American, African unspecified, European
Lung function forced vital capacity	-11.34752	8.99E-03	Chocó	East Asian, Hispanic or Latin American, European, African American or Afro-Caribbean

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Lung function FVC	4.0658476	6.13E-04	Antioquia	East Asian, Hispanic or Latin American, African unspecified, European
Lymphoma	-11.34752	7.22E-04	Chocó	NR, European
Mammographic density	-8.191489	1.73E-02	Chocó	European
Mammographic density dense area	-5.817215	3.65E-03	Chocó	NR, European, Other
Maximum cranial width	-6.540853	1.29E-06	Chocó	European
Mean arterial pressure	2.2651123	1.25E-02	Antioquia	East Asian
Mean arterial pressure x alcohol consumption light vs heavy interaction 2df test	5.1685888	1.81E-04	Antioquia	African American or Afro-Caribbean, Asian unspecified, European, African unspecified, Hispanic or Latin American
Mean corpuscular hemoglobin	2.2065264	9.18E-11	Antioquia	East Asian
Mean corpuscular hemoglobin concentration	3.7266113	5.42E-06	Antioquia	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Mean corpuscular volume	3.0844077	1.67E-17	Antioquia	East Asian
Mean platelet volume	1.6620452	2.85E-07	Antioquia	European
Melanoma	6.7520954	9.07E-05	Antioquia	European
Membranous nephropathy	-15.15957	2.29E-05	Chocó	European
Memory dysfunction in frontotemporal lobe dementia	-13.12943	4.62E-08	Chocó	European
Menarche and menopause age at onset	-16.17021	6.04E-11	Chocó	European
Menopause age at onset	-4.186097	6.33E-08	Chocó	East Asian, European
Menstruation quality of life impact acne	-8.118034	2.66E-03	Chocó	East Asian
Menstruation quality of life impact headache	-10.90426	1.93E-02	Chocó	East Asian
Metabolite levels	-5.167105	3.93E-08	Chocó	European
Metabolite levels X-11787	4.076837	1.50E-02	Antioquia	African American or Afro-Caribbean
Metabolite levels HVA-5-HIAA Factor score	-15.98727	1.01E-21	Chocó	European
Methotrexate phramacokinetics acute lymphoblastic leukemia	9.2404614	3.36E-20	Antioquia	African American or Afro-Caribbean, Asian unspecified, European, Other, NR

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
MGMT methylation in smokers	-20.47872	4.52E-07	Chocó	European, Hispanic or Latin American
Migraine - clinic-based	-7.31383	1.67E-07	Chocó	European
Migraine with aura	-13.78143	4.09E-15	Chocó	European
Monocyte chemoattractant protein-1 levels	-3.035025	8.33E-03	Chocó	European
Monocyte chemoattractant protein-3 levels	5.3698075	1.60E-02	Antioquia	European
Monocyte count	1.6290198	1.02E-04	Antioquia	East Asian
Monocyte percentage of white cells	3.2923709	5.83E-16	Antioquia	European
Mortality in heart failure	-12.14096	1.27E-11	Chocó	European
Mortality in sepsis	-8.251224	7.84E-07	Chocó	NR, European
Multiple myeloma	-4.326545	1.13E-04	Chocó	European
Multiple sclerosis	-1.154707	3.28E-02	Chocó	European
Multiple system atrophy	-5.649696	2.03E-03	Chocó	European
Myeloproliferative neoplasms	-16.48936	5.41E-07	Chocó	European
Myocardial infarction early onset	7.5132979	3.28E-04	Antioquia	European
Nasopharyngeal carcinoma	6.8085106	9.01E-05	Antioquia	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Neonatal cytokine-over-chemokine levels fetal genetic effect	-7.255784	2.15E-05	Chocó	European, South Asian, Asian unspecified, African American or Afro-Caribbean, Hispanic or Latin American
Neovascular age-related macular degeneration	12.721631	5.14E-07	Antioquia	South East Asian
Neuroblastoma 1p deletion	-26.06383	7.63E-05	Chocó	European
Neurociticism	1.5732397	3.85E-02	Antioquia	European
Neurocognitive impairment in HIV-1 infection continuous	5.6117021	1.41E-03	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American, NR
Neurofibrillary tangles	-6.095775	2.07E-03	Chocó	NR
Neuroticism	1.382575	1.36E-06	Antioquia	European
Neutrophil count	2.4585936	2.78E-06	Antioquia	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Neutrophil count in HIV-infection	-73.93617	9.44E-45	Chocó	European, Hispanic or Latin American, African unspecified
Neutrophil level response to clozapine in treatment-resistant schizophrenia	-73.93617	9.44E-45	Chocó	Sub-Saharan African
Neutrophil percentage of granulocytes	3.5444048	4.59E-12	Antioquia	European
Neutrophil percentage of white cells	2.4744523	2.46E-05	Antioquia	European
Nicotine use	-8.244681	1.14E-05	Chocó	European
Non-albumin protein levels	-27.83688	5.47E-15	Chocó	East Asian
Nonalcoholic fatty liver disease	9.5803783	5.62E-04	Antioquia	East Asian
Non-alcoholic fatty liver disease histology other	6.1057659	5.09E-10	Antioquia	European
Non-melanoma skin cancer	-3.10881	1.36E-03	Chocó	NR, European
Non-small cell lung cancer survival	12.234043	5.27E-04	Antioquia	East Asian
Number of children ever born	14.361702	2.59E-03	Antioquia	European
Obesity early onset extreme	-10.06891	1.76E-15	Chocó	European
Obesity-related traits	-2.911325	1.73E-39	Chocó	Hispanic or Latin American

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Obstructive sleep apnea during REM sleep apnea hypopnea index	6.4716312	4.27E-03	Antioquia	East Asian, African American or Afro-Caribbean, Asian unspecified, European, Hispanic or Latin American
Obstructive sleep apnea trait apnea hypopnea index	6.8955513	1.16E-08	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American, Asian unspecified
Obstructive sleep apnea trait average respiratory event duration	8.5106383	1.42E-03	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American, Asian unspecified
Oppositional defiant disorder dimensions in attention-deficit hyperactivity disorder	-6.916164	7.79E-04	Chocó	NR, European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Optic disc area	-30.67059	7.24E-32	Chocó	South East Asian, South Asian, East Asian, European
Optic nerve measurement rim area	9.0248227	3.10E-04	Antioquia	European
Ossification of the posterior longitudinal ligament of the spine	-5.868794	1.74E-02	Chocó	East Asian
Osteoarthritis of the hip with total joint replacement	-13.93617	2.02E-06	Chocó	European
Osteoarthritis of the knee hospital diagnosed	7.2251773	7.01E-03	Antioquia	European
Osteoporosis-related phenotypes	8.9184397	2.10E-04	Antioquia	European
Osteoprotegerin levels	-12.05674	1.96E-02	Chocó	East Asian, European
Overall survival in osteosarcoma	-11.28314	5.28E-11	Chocó	European, Hispanic or Latin American
Overweight status	-14.09574	7.69E-05	Chocó	South Asian
Pain	21.808511	2.99E-03	Antioquia	European
Pain medicine use during menstruation	-7.593718	1.30E-02	Chocó	East Asian
Palmitic acid 16:0 levels	8.7505373	4.68E-08	Antioquia	European
Pancreatitis	15.602837	2.73E-05	Antioquia	European
Paraoxonase activity	-25	1.10E-06	Chocó	European
Parental longevity combined parental age at death	11.365248	4.49E-08	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Parental longevity combined parental attained age, Martingale residuals	-5.485727	2.39E-06	Chocó	European
Parental longevity father's age at death	7.3138298	3.98E-02	Antioquia	European
Parkinsonism in frontotemporal lobe dementia	12.85461	8.18E-04	Antioquia	European
Parkinson's disease	2.8948549	1.84E-04	Antioquia	European
Parkinson's disease age of onset	-19.90248	1.78E-11	Chocó	NR, European
Pediatric autoimmune diseases	7.3415744	7.04E-14	Antioquia	European
Pediatric bone mineral content femoral neck	7.385385	1.29E-02	Antioquia	European
Pediatric bone mineral content hip	-5.217114	1.28E-02	Chocó	European
Pediatric bone mineral density femoral neck	5.9553656	1.13E-03	Antioquia	European
Pediatric bone mineral density hip	9.2907801	2.11E-08	Antioquia	European
Perceived skin darkness	-22.07447	1.55E-07	Chocó	European
Percent glycated albumin	11.170213	1.42E-02	Antioquia	African unspecified, European
Percent mammographic density	-10.74468	7.02E-04	Chocó	NR, European, Other
Periodontitis DPAL	6.6986353	2.00E-04	Antioquia	European
Periodontitis PAL4Q3	11.409786	4.30E-09	Antioquia	European
Pharmacokinetics of antidepressant drugs in severe mental disorder concentration dose ratio	-26.06383	3.09E-05	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Philtrum width	-23.67021	6.92E-10	Chocó	NR, European
Photic sneeze reflex	-8.599291	2.33E-03	Chocó	European
Physical activity moderate intensity activity duration	16.223404	2.98E-03	Antioquia	European
Pit-and-Fissure caries	5.4477437	1.67E-02	Antioquia	European
Placental abruption	-8.076241	1.56E-04	Chocó	NR
Plasma amyloid beta peptide concentrations ABx-40	-10.63197	5.54E-07	Chocó	European
Plasma homocysteine levels post-methionine load test	15.691489	1.13E-06	Antioquia	African unspecified, European, Other, NR, European
Plasma neurofilament light levels	-22.87234	1.34E-02	Chocó	European
Plasma omega-3 polyunsaturated fatty acid level eicosapentaenoic acid	4.4232216	3.07E-03	Antioquia	European
Plasma omega-3 polyunsaturated fatty acid levels docosahexaenoic acid	5.5972839	5.38E-04	Antioquia	East Asian
Plasma trimethyllysine levels	9.751773	1.33E-04	Antioquia	European
Platelet count	2.0467445	3.35E-11	Antioquia	East Asian
Platelet distribution width	2.7499875	3.06E-11	Antioquia	European
Plateletcrit	3.914265	2.37E-20	Antioquia	European
Platelet-derived growth factor BB levels	-6.338909	4.21E-07	Chocó	European
Polycystic ovary syndrome	-5.086455	3.21E-05	Chocó	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Post bronchodilator FEV1	1.6890387	2.94E-12	Antioquia	African American or Afro-Caribbean, European, NR
Post bronchodilator FEV1-over-FVC ratio in smoking	10.638298	3.82E-11	Antioquia	European
Post-traumatic stress disorder	12.748227	3.85E-05	Antioquia	European
PR segment	12.319529	4.82E-09	Antioquia	European
Preschool internalizing problems	6.0345476	1.13E-05	Antioquia	NR, European
Prevalent atrial fibrillation	9.9037487	1.77E-07	Antioquia	East Asian, European, NR, African American or Afro-Caribbean, Hispanic or Latin American
Primary tooth development number of teeth	6.8345849	2.12E-11	Antioquia	European
Progression free survival in metastatic colorectal cancer CAPOX-B vs CAPOX-B plus cetuximab	-19.32624	2.18E-10	Chocó	European
Proliferative diabetic retinopathy in type 2 diabetes	-27.12766	1.94E-04	Chocó	European
Prostate cancer advanced	7.8116269	3.19E-06	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Prostate-specific antigen levels	-3.893027	2.99E-03	Chocó	East Asian, European, Hispanic or Latin American, African American or Afro-Caribbean
Proteinuria and chronic kidney disease	11.968085	1.08E-03	Antioquia	European, Other
Psoriasis	3.0482491	3.74E-04	Antioquia	European
Psoriatic arthritis	4.8063904	4.22E-02	Antioquia	European
Pulmonary emphysema	15.248227	6.72E-07	Antioquia	East Asian, European, Hispanic or Latin American, African American or Afro-Caribbean
Pulmonary function in asthmatics	-9.734887	2.80E-08	Chocó	European
Pulse pressure	1.2860061	3.38E-02	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Pulse pressure x alcohol consumption interaction 2df test	3.3118134	5.67E-04	Antioquia	African American or Afro-Caribbean, Asian unspecified, European, African unspecified, Hispanic or Latin American
QRS duration	5.1342413	4.48E-04	Antioquia	African American or Afro-Caribbean, European, South Asian
QT interval drug interaction	-10.74673	1.07E-10	Chocó	European
RANTES levels	3.4406905	1.13E-02	Antioquia	European
Ratio of the peak velocity of the mitral E-Wave divided by the peak velocity of the mitral A-wave.	-46.2766	3.75E-15	Chocó	European
Recalcitrant atopic dermatitis	-8.271277	1.50E-02	Chocó	East Asian
Recurrent major depressive disorder	4.6564163	3.01E-02	Antioquia	European
Red blood cell count	2.7376659	1.74E-10	Antioquia	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Red blood cell density in sickle cell anemia	-5.845576	1.56E-03	Chocó	NR, African unspecified, African unspecified
Red cell distribution width	3.3158002	1.00E-13	Antioquia	European
Regular attendance at a religious group	-9.043422	1.23E-11	Chocó	European
Remission after SSRI treatment in MDD or neuroticism	-10.47366	1.18E-05	Chocó	NR
Renal function-related traits BUN	9.0251947	8.81E-08	Antioquia	East Asian, South Asian, South East Asian
Renal function-related traits eGRFcrea	7.8457447	1.60E-03	Antioquia	East Asian, South Asian, South East Asian
Renal function-related traits urea	-19.41489	7.86E-04	Chocó	East Asian, South Asian, South East Asian
Renal underexcretion gout	-5.878926	2.13E-02	Chocó	East Asian, European, Oceanian
Residual cognition	-7.045762	8.98E-06	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Resistance to antihypertensive treatment in hypertension	24.468085	4.01E-05	Antioquia	European, African American or Afro-Caribbean, Hispanic or Latin American, NR
Resistin levels	-27.34929	7.45E-20	Chocó	East Asian
Response to anti-depressant treatment in major depressive disorder	-2.77741	2.05E-03	Chocó	African American or Afro-Caribbean, European
Response to anti-TNF therapy in rheumatoid arthritis	-6.458333	5.36E-03	Chocó	East Asian
Response to gemcitabine in pancreatic cancer	-9.911348	7.50E-09	Chocó	East Asian
Response to lurasidone in schizophrenia	-13.9273	6.35E-10	Chocó	African American or Afro-Caribbean, European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Response to metformin in type 2 diabetes HbA1c reduction	-28.19149	1.11E-04	Chocó	Asian unspecified, European, Hispanic or Latin American, African American or Afro-Caribbean
Response to methotrexate in rheumatoid arthritis	-14.09574	5.76E-05	Chocó	South Asian
Response to methylphenidate treatment in attention-deficit-over-hyperactivity disorder blood pressure	-18.49291	5.60E-23	Chocó	African American or Afro-Caribbean, Hispanic or Latin American, Other
Response to montelukast in asthma change in FEV1	-16.75532	2.21E-02	Chocó	Asian unspecified, African unspecified, European
Response to olanzapine in schizophrenia	-15.15957	1.36E-02	Chocó	East Asian
Response to perphenazine in schizophrenia	-15.69149	2.48E-08	Chocó	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Response to simvastatin treatment PCSK9 protein level change	-8.81459	3.26E-05	Chocó	African American or Afro-Caribbean, European
Response to SSRI in MDD or openness	-13.1383	2.82E-10	Chocó	NR
Response to statin therapy	-3.337924	1.08E-03	Chocó	European
Response to tamoxifen in breast cancer	30.851064	2.58E-07	Antioquia	East Asian
Response to Vitamin E supplementation	9.3085106	1.32E-02	Antioquia	European
Resting-state electroencephalogram vigilance	-3.537472	1.17E-04	Chocó	European
Reticulocyte count	1.9674487	2.89E-05	Antioquia	European
Reticulocyte fraction of red cells	1.8814177	5.45E-05	Antioquia	European
Retinal arteriolar caliber	-14.62766	1.23E-03	Chocó	European
Retinol levels	-17.81915	7.01E-04	Chocó	NR, European
Rhegmatogenous retinal detachment	-10.19504	9.86E-12	Chocó	European
Rosacea symptom severity	4.0255965	5.96E-10	Antioquia	European
Schizophrenia	-2.220293	1.18E-15	Chocó	NR, European
Self-rated health	-9.413416	5.06E-06	Chocó	European
Self-reported allergy	5.8550646	2.99E-07	Antioquia	European
Self-reported math ability	-3.958835	8.89E-41	Chocó	European
Self-reported math ability MTAG	1.2235144	1.11E-08	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Serum 25-Hydroxyvitamin D levels	17.408604	2.79E-14	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American
Serum albumin level	-14.89362	1.51E-03	Chocó	East Asian, European
Serum C3d:C3 ratio systemic complement activation	14.62766	1.51E-07	Antioquia	European
Serum ceruloplasmin levels	-25.53191	1.05E-05	Chocó	European
Serum dimethylarginine levels asymmetric-over-symetric ratio	-7.730496	2.26E-02	Chocó	European
Serum tamsulosin hydrochloride concentration	8.9539007	2.32E-04	Antioquia	East Asian
Serum urate levels in chronic kidney disease	15.966312	9.54E-13	Antioquia	European
Serum vitamin D-binding protein levels	5.6737589	4.22E-02	Antioquia	European
Severity of facial solar lentigines	8.2446809	4.09E-02	Antioquia	European
Severity of nausea and vomiting of pregnancy	7.7039007	1.84E-02	Antioquia	European
Sex hormone-binding globulin levels	-6.380189	4.33E-04	Chocó	European
Sexual dysfunction SSRI-over-SNRI-related	5.106383	4.65E-06	Antioquia	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Sitting height ratio	-4.403614	1.88E-04	Chocó	African American or Afro-Caribbean, European
Sjögren's syndrome	4.842304	1.59E-03	Antioquia	European
Skin colour saturation	-34.17553	7.70E-33	Chocó	European
Skin pigmentation	-8.156178	1.35E-20	Chocó	Sub-Saharan African
Skin sensitivity to sun	-57.44681	2.08E-24	Chocó	European
Sleep duration	-6.906329	5.76E-12	Chocó	European
Sleep quality	-11.09929	2.12E-06	Chocó	European
Smoking cessation in chronic obstructive pulmonary disease	-9.014691	1.43E-04	Chocó	European
Smoking status heavy vs light	23.404255	4.38E-04	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American
Smooth-surface caries	4.4678668	3.87E-09	Antioquia	European
Social communication problems	-4.324106	3.63E-02	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Glomerular filtration rate change in heart transplantation	0.24	1.0E+00	Antioquia	NR, European, Native American
Glomerular filtration rate creatinine	4.42	3.7E-17	Antioquia	European
Glomerular filtration rate cystatin C	-2.13	1.0E+00	Chocó	European
Glomerular filtration rate in chronic kidney disease	2.46	1.0E+00	Antioquia	European, Other
Glomerular filtration rate in non diabetics creatinine	5.44	9.7E-07	Antioquia	Asian unspecified, African unspecified, European
GLP-1 levels in response to oral glucose tolerance test 120 minutes	1.06	1.0E+00	Antioquia	European
Glucocorticoid-induced osteonecrosis	-3.89	1.0E+00	Chocó	East Asian, African American or Afro-Caribbean, European, NR, Hispanic or Latin American
Glycated hemoglobin levels	1.48	1.0E+00	Antioquia	East Asian
Glycemic traits	0.00	NA	Antioquia	East Asian
Glycemic traits pregnancy	23.85	6.7E-18	Antioquia	African American or Afro-Caribbean, European, Hispanic or Latin American, South East Asian
Glycerophospholipid levels	-7.62	1.0E+00	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Spherical equivalent	5.8388537	1.46E-03	Antioquia	South Asian, East Asian, South East Asian, Asian unspecified, European
Sporadic neuroblastoma	-14.98227	5.64E-12	Chocó	African American or Afro-Caribbean, European, NR
Squamous cell lung carcinoma	3.0243053	9.23E-07	Antioquia	European
Stearic acid 18:0 levels	10.437774	1.19E-08	Antioquia	European
Stroke	3.6815139	1.56E-02	Antioquia	East Asian, South Asian, African American or Afro-Caribbean, Asian unspecified, European, Hispanic or Latin American
Stroke ischemic	13.244681	6.04E-10	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Subclinical trait of interstitial lung disease basilar peel-core ratio of high attentuation areas on CT scan	-5.806738	2.97E-03	Chocó	East Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Subclinical trait of interstitial lung disease basilar percentage of high attenuation areas on CT scan	-5.585106	1.58E-02	Chocó	East Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Subcortical brain region volumes	13.297872	1.21E-09	Antioquia	East Asian, European, Hispanic or Latin American
Sub-foveal choroidal thickness	-39.89362	1.20E-10	Chocó	East Asian
Subiculum volume	12.721631	9.33E-06	Antioquia	European
Subjective response to lithium treatment in bipolar disorder	11.746454	2.20E-03	Antioquia	European
Suicide attempts	9.5110176	6.50E-07	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Suicide ideation score in major depressive disorder	-13.65248	5.81E-06	Chocó	European
Suicide in bipolar disorder	-15.0266	4.52E-05	Chocó	European
Sum basophil neutrophil counts	2.6969014	1.13E-07	Antioquia	European
Supraventricular ectopy	-21.03723	6.31E-22	Chocó	African unspecified, Hispanic or Latin American, European
Survival in colorectal cancer distant metastatic	-15.15957	9.53E-04	Chocó	European
Systemic lupus erythematosus	-3.844557	8.56E-25	Chocó	European
Systolic blood pressure cigarette smoking interaction	-8.619554	2.61E-06	Chocó	East Asian, South Asian, African American or Afro-Caribbean, Sub-Saharan African, Asian unspecified, European, Hispanic or Latin American

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Systolic blood pressure in sickle cell anemia	9.0984887	3.89E-07	Antioquia	African American or Afro-Caribbean
Systolic blood pressure response to hydrochlorothiazide in hypertension	6.6770095	7.30E-06	Antioquia	African American or Afro-Caribbean, European
Systolic blood pressure x alcohol consumption interaction 2df test	2.4788497	8.95E-04	Antioquia	African American or Afro-Caribbean, Asian unspecified, European, African unspecified, Hispanic or Latin American

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Systolic blood pressure x alcohol consumption light vs heavy interaction 2df test	-5.187215	3.65E-07	Chocó	African American or Afro-Caribbean, Asian unspecified, European, African unspecified, Hispanic or Latin American
Testicular cancer	26.41844	7.01E-13	Antioquia	European
Thiazide-induced adverse metabolic effects in hypertensive patients	3.8976321	9.91E-08	Antioquia	African American or Afro-Caribbean, European
Thrombin-activatable fibrinolysis inhibitor activation peptide	-10.8156	9.17E-05	Chocó	European
Thyroid cancer Papillary, radiation-related	17.553191	2.70E-02	Antioquia	NR, European
Thyroid function	12.411348	2.25E-02	Antioquia	European
Thyroid hormone levels	22.340426	4.60E-05	Antioquia	European
Thyroid peroxidase antibody levels	15.957447	1.36E-09	Antioquia	European
Thyroid peroxidase antibody positivity	8.4736998	5.50E-04	Antioquia	East Asian
Thyroid peroxidase autoantibody levels in type 1 diabetes	26.595745	4.88E-05	Antioquia	NR, European
Thyroid stimulating hormone	-5.816594	6.33E-05	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Tooth agenesis mandibular second premolars	-35.10638	6.88E-12	Chocó	European, NR
Total body bone mineral density	2.4686377	8.46E-05	Antioquia	European
Total body bone mineral density age 45-60	4.9336568	4.48E-02	Antioquia	European, NR, African American or Afro-Caribbean
Total body bone mineral density age over 60	6.0935917	2.83E-05	Antioquia	European, NR, African American or Afro-Caribbean
Total cholesterol levels	-2.11856	2.74E-06	Chocó	African American or Afro-Caribbean, European, Hispanic or Latin American, NR
Tourette's syndrome or obsessive-compulsive disorder	-8.228128	2.07E-07	Chocó	European
Triglycerides-Blood Pressure TG-BP	7.2695035	1.80E-05	Antioquia	European
Tuberculosis	4.548807	3.51E-05	Antioquia	East Asian
Tumor biomarkers	-14.04509	3.76E-14	Chocó	East Asian

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Tumor necrosis factor alpha levels	-5.871327	4.21E-03	Chocó	European
Two-hour glucose challenge	10.585106	3.86E-04	Antioquia	European, European, Other
Type 1 diabetes	-3.887702	2.26E-05	Chocó	European
Type 2 diabetes	-1.796435	2.40E-05	Chocó	European
Type 2 diabetes and other traits	-7.624113	1.06E-02	Chocó	European
Type 2 diabetes nephropathy	-23.7766	4.42E-23	Chocó	East Asian
Upper eyelid morphology	7.8723404	4.72E-03	Antioquia	East Asian
Upper eyelid sagging severity	-4.205164	1.37E-02	Chocó	European
Urate levels BMI interaction	4.258304	9.15E-03	Antioquia	European
Urea levels	18.218085	6.44E-12	Antioquia	European
Urinary albumin-to-creatinine ratio	8.4542231	7.28E-08	Antioquia	Hispanic or Latin American
Urinary electrolytes magnesium-over-calcium ratio	-18.70567	9.22E-08	Chocó	European
Urinary metabolite ratios in chronic kidney disease	-14.22872	1.62E-02	Chocó	European
Urinary tract infection frequency	-17.55319	5.93E-09	Chocó	European
Vascular endothelial growth factor levels	5.4204783	1.05E-07	Antioquia	European
Verbal memory performance residualized delayed recall change	5.5851064	2.13E-04	Antioquia	European
Verbal memory performance residualized delayed recall level	-10.90679	1.19E-09	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Vertical cup-disc ratio	-7.88584	1.37E-06	Chocó	South East Asian, South Asian, East Asian, European
Very long-chain saturated fatty acid levels fatty acid 20:0	5.3350904	9.35E-03	Antioquia	European
Visceral adipose tissue	-12.76596	1.89E-07	Chocó	East Asian, African American or Afro-Caribbean, Asian unspecified, European, Hispanic or Latin American
Visceral fat	-3.272413	1.35E-02	Chocó	European
Vitamin B12 levels	-7.624113	5.76E-09	Chocó	South Asian
Vitamin D levels dietary vitamin D intake interaction	-16.0461	1.79E-07	Chocó	European, NR
Vitamin E levels	12.234043	2.36E-03	Antioquia	European, Other
Vogt-Koyanagi-Harada syndrome	15.425532	1.10E-02	Antioquia	East Asian
vWF and FVIII levels	-6.576402	2.64E-06	Chocó	European
Waist Circumference - Triglycerides WC-TG	-18.9273	3.07E-11	Chocó	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Waist circumference adjusted for BMI in non-smokers	4.743928	2.99E-06	Antioquia	East Asian, South Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Waist circumference adjusted for BMI joint analysis main effects and smoking interaction	4.1804563	2.97E-04	Antioquia	East Asian, South Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Waist-hip ratio	5.6317353	1.53E-05	Antioquia	European
Waist-to-hip ratio adjusted for BMI	4.3682369	9.38E-13	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
Waist-to-hip ratio adjusted for BMI in non-smokers	5.5371297	2.10E-05	Antioquia	East Asian, South Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Waist-to-hip ratio adjusted for BMI x sex interaction	8.7242438	9.38E-14	Antioquia	European
Waist-to-hip ratio adjusted for BMI x sex x age interaction 4df test	5.0693275	3.11E-15	Antioquia	European
White blood cell count	3.4422366	1.16E-15	Antioquia	East Asian
White blood cell count monocyte	16.115755	5.55E-19	Antioquia	Hispanic or Latin American
White blood cell count neutrophil	10.01773	7.56E-07	Antioquia	Hispanic or Latin American
White matter hyperintensities in ischemic stroke	10	4.90E-06	Antioquia	European

Table 7 continued

GWAS Catalog Trait	ΔPTS	Holm-Bonferroni P-value	Higher In Population	Study population(s)
White matter hyperintensity burden	-8.235605	5.96E-03	Chocó	East Asian, South East Asian, African American or Afro-Caribbean, European, Hispanic or Latin American
Wilms tumor	-18.19149	5.01E-22	Chocó	European
Worry too long after an embarrassing experience	-5.844995	2.02E-08	Chocó	European
Yeast infection	-15.15957	3.28E-03	Chocó	European

Table 8. PRS derived from ancestry-specific studies

rsid	EA	effect_size	Trait	PMID	Accession
rs11210537	A	5.824	Myopia-Multiethnic	29808027	GCST006291
rs11589487	A	6.665	Myopia-Multiethnic	29808027	GCST006291
rs1237670	A	5.815	Myopia-Multiethnic	29808027	GCST006291
rs11802995	A	6.203	Myopia-Multiethnic	29808027	GCST006291
rs1556867	T	8.806	Myopia-Multiethnic	29808027	GCST006291
rs2225986	A	7.963	Myopia-Multiethnic	29808027	GCST006291
rs1858001	C	7.276	Myopia-Multiethnic	29808027	GCST006291
rs2745953	A	5.759	Myopia-Multiethnic	29808027	GCST006291
rs11118367	T	7.288	Myopia-Multiethnic	29808027	GCST006291
rs6753137	T	6.39E+00	Myopia-Multiethnic	29808027	GCST006291
rs28658452	A	5.49	Myopia-Multiethnic	29808027	GCST006291
rs17032696	A	5.766	Myopia-Multiethnic	29808027	GCST006291
rs41393947	A	6.42E+00	Myopia-Multiethnic	29808027	GCST006291
rs10187371	T	6.561	Myopia-Multiethnic	29808027	GCST006291
rs56075542	T	8.994	Myopia-Multiethnic	29808027	GCST006291
rs297593	T	7.816	Myopia-Multiethnic	29808027	GCST006291
rs17428076	C	8.183	Myopia-Multiethnic	29808027	GCST006291
rs2573081	C	8.212	Myopia-Multiethnic	29808027	GCST006291
rs1550094	A	12.738	Myopia-Multiethnic	29808027	GCST006291
rs12998513	A	6.855	Myopia-Multiethnic	29808027	GCST006291
rs931302	T	5.75E+00	Myopia-Multiethnic	29808027	GCST006291
rs9681162	T	6.699	Myopia-Multiethnic	29808027	GCST006291
rs1454776	T	6.65	Myopia-Multiethnic	29808027	GCST006291
rs4260345	T	5.662	Myopia-Multiethnic	29808027	GCST006291
rs4687586	C	6.547	Myopia-Multiethnic	29808027	GCST006291
rs2303635	C	6.088	Myopia-Multiethnic	29808027	GCST006291
rs7624084	T	8.811	Myopia-Multiethnic	29808027	GCST006291
rs4894529	A	5.646	Myopia-Multiethnic	29808027	GCST006291
rs7662551	A	8.53	Myopia-Multiethnic	29808027	GCST006291
rs7747	T	7.031	Myopia-Multiethnic	29808027	GCST006291
rs1994840	A	5.566	Myopia-Multiethnic	29808027	GCST006291
rs7692381	A	9.398	Myopia-Multiethnic	29808027	GCST006291
rs2166181	A	6.642	Myopia-Multiethnic	29808027	GCST006291
rs11952819	T	5.795	Myopia-Multiethnic	29808027	GCST006291
rs7737179	A	5.527	Myopia-Multiethnic	29808027	GCST006291
rs7449443	T	5.51	Myopia-Multiethnic	29808027	GCST006291
rs10458138	A	6.049	Myopia-Multiethnic	29808027	GCST006291
rs144370238	T	6.046	Myopia-Multiethnic	29808027	GCST006291

Table 8 comtinued

rsid	EA	effect_size	Trait	PMID	Accession
rs1207782	T	7.915	Myopia-Multiethnic	29808027	GCST006291
rs1150687	T	6.784	Myopia-Multiethnic	29808027	GCST006291
rs9395623	A	7.25	Myopia-Multiethnic	29808027	GCST006291
rs7744813	A	14.555	Myopia-Multiethnic	29808027	GCST006291
rs1064583	A	6.52E+00	Myopia-Multiethnic	29808027	GCST006291
rs12193446	A	19.431	Myopia-Multiethnic	29808027	GCST006291
rs9388766	T	6.23E+00	Myopia-Multiethnic	29808027	GCST006291
rs1358684	T	5.512	Myopia-Multiethnic	29808027	GCST006291
rs12667032	A	7.994	Myopia-Multiethnic	29808027	GCST006291
rs60884546	A	6.23E+00	Myopia-Multiethnic	29808027	GCST006291
rs2116093	C	5.527	Myopia-Multiethnic	29808027	GCST006291
rs10104039	T	5.582	Myopia-Multiethnic	29808027	GCST006291
rs1532278	T	5.69E+00	Myopia-Multiethnic	29808027	GCST006291
rs7829127	A	10.911	Myopia-Multiethnic	29808027	GCST006291
rs284816	A	7.212	Myopia-Multiethnic	29808027	GCST006291
rs72621438	C	13.137	Myopia-Multiethnic	29808027	GCST006291
rs72655575	A	6.867	Myopia-Multiethnic	29808027	GCST006291
rs28891973	A	6.27E+00	Myopia-Multiethnic	29808027	GCST006291
rs10511652	A	7.355	Myopia-Multiethnic	29808027	GCST006291
rs11145465	A	9.546	Myopia-Multiethnic	29808027	GCST006291
rs7042950	A	6.797	Myopia-Multiethnic	29808027	GCST006291
rs10760673	A	5.986	Myopia-Multiethnic	29808027	GCST006291
rs10122788	A	5.491	Myopia-Multiethnic	29808027	GCST006291
rs11101263	T	7.329	Myopia-Multiethnic	29808027	GCST006291
rs1649068	A	9.439	Myopia-Multiethnic	29808027	GCST006291
rs9416017	T	5.655	Myopia-Multiethnic	29808027	GCST006291
rs4237284	A	5.476	Myopia-Multiethnic	29808027	GCST006291
rs7895108	T	8.866	Myopia-Multiethnic	29808027	GCST006291
rs745480	C	8.314	Myopia-Multiethnic	29808027	GCST006291
rs11202736	A	6.919	Myopia-Multiethnic	29808027	GCST006291
rs17382981	T	6.31E+00	Myopia-Multiethnic	29808027	GCST006291
rs807037	C	5.466	Myopia-Multiethnic	29808027	GCST006291
rs72826094	A	7.883	Myopia-Multiethnic	29808027	GCST006291
rs511217	A	6.793	Myopia-Multiethnic	29808027	GCST006291
rs7941828	T	5.483	Myopia-Multiethnic	29808027	GCST006291
rs11602008	A	13.978	Myopia-Multiethnic	29808027	GCST006291
rs7107014	A	5.459	Myopia-Multiethnic	29808027	GCST006291
rs478304	T	6.099	Myopia-Multiethnic	29808027	GCST006291

Table 8 comtinued

rsid	EA	effect_size	Trait	PMID	Accession
rs2155413	A	7.755	Myopia-Multiethnic	29808027	GCST006291
rs1954761	T	8.397	Myopia-Multiethnic	29808027	GCST006291
rs7122817	A	7.514	Myopia-Multiethnic	29808027	GCST006291
rs1790165	A	6.854	Myopia-Multiethnic	29808027	GCST006291
rs4764038	T	6.196	Myopia-Multiethnic	29808027	GCST006291
rs7971334	T	5.724	Myopia-Multiethnic	29808027	GCST006291
rs117735470	A	6.07	Myopia-Multiethnic	29808027	GCST006291
rs10880855	T	7.778	Myopia-Multiethnic	29808027	GCST006291
rs3138141	A	13.803	Myopia-Multiethnic	29808027	GCST006291
rs11178469	T	7.403	Myopia-Multiethnic	29808027	GCST006291
rs7337610	T	6.144	Myopia-Multiethnic	29808027	GCST006291
rs1359543	A	6.30E+00	Myopia-Multiethnic	29808027	GCST006291
rs9547035	T	5.69E+00	Myopia-Multiethnic	29808027	GCST006291
rs9516194	A	6.25E+00	Myopia-Multiethnic	29808027	GCST006291
rs9517964	T	8.423	Myopia-Multiethnic	29808027	GCST006291
rs837323	T	6.32E+00	Myopia-Multiethnic	29808027	GCST006291
rs12883788	T	6.24E+00	Myopia-Multiethnic	29808027	GCST006291
rs36024104	A	9.09	Myopia-Multiethnic	29808027	GCST006291
rs2855530	C	8.575	Myopia-Multiethnic	29808027	GCST006291
rs2753462	C	6.49E+00	Myopia-Multiethnic	29808027	GCST006291
rs17125093	A	6.23E+00	Myopia-Multiethnic	29808027	GCST006291
rs11160044	A	6.531	Myopia-Multiethnic	29808027	GCST006291
rs35337422	A	6.35E+00	Myopia-Multiethnic	29808027	GCST006291
rs524952	A	17.075	Myopia-Multiethnic	29808027	GCST006291
rs34539187	C	5.808	Myopia-Multiethnic	29808027	GCST006291
rs12898755	A	7.533	Myopia-Multiethnic	29808027	GCST006291
rs6495367	A	10.202	Myopia-Multiethnic	29808027	GCST006291
rs1969091	A	6.37E+00	Myopia-Multiethnic	29808027	GCST006291
rs79266634	C	5.932	Myopia-Multiethnic	29808027	GCST006291
rs10500355	A	13.732	Myopia-Multiethnic	29808027	GCST006291
rs56055503	A	6.719	Myopia-Multiethnic	29808027	GCST006291
rs8075280	A	5.75E+00	Myopia-Multiethnic	29808027	GCST006291
rs2908972	A	11.125	Myopia-Multiethnic	29808027	GCST006291
rs62070229	A	8.578	Myopia-Multiethnic	29808027	GCST006291
rs3213636	A	6.26E+00	Myopia-Multiethnic	29808027	GCST006291
rs4795364	A	5.532	Myopia-Multiethnic	29808027	GCST006291
rs11654644	T	6.32E+00	Myopia-Multiethnic	29808027	GCST006291
rs12451582	A	7.02	Myopia-Multiethnic	29808027	GCST006291

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs4793501	T	7.212	Myopia-Multiethnic	29808027	GCST006291
rs6420484	A	5.943	Myopia-Multiethnic	29808027	GCST006291
rs10853531	A	6.882	Myopia-Multiethnic	29808027	GCST006291
rs12965607	T	7.073	Myopia-Multiethnic	29808027	GCST006291
rs4808962	A	6.056	Myopia-Multiethnic	29808027	GCST006291
rs4805962	T	5.773	Myopia-Multiethnic	29808027	GCST006291
rs235770	T	5.926	Myopia-Multiethnic	29808027	GCST006291
rs1555075	T	6.215	Myopia-Multiethnic	29808027	GCST006291
rs2229742	C	6.121	Myopia-Multiethnic	29808027	GCST006291
rs11088317	T	6.895	Myopia-Multiethnic	29808027	GCST006291
rs9680365	A	5.751	Myopia-Multiethnic	29808027	GCST006291
rs2150458	A	7.735	Myopia-Multiethnic	29808027	GCST006291
rs9606967	C	5.866	Myopia-Multiethnic	29808027	GCST006291
rs1983554	A	5.51	Myopia-Multiethnic	29808027	GCST006291
rs10910076	A	0.648	Myopia-European	29808027	GCST006290
rs479445	A	0.0481	Myopia-European	29808027	GCST006290
rs57382675	A	0.0641	Myopia-European	29808027	GCST006290
rs10919908	A	0.0507	Myopia-European	29808027	GCST006290
rs2162488	T	0.1078	Myopia-European	29808027	GCST006290
rs61049169	A	0.0629	Myopia-European	29808027	GCST006290
rs297587	A	0.0497	Myopia-European	29808027	GCST006290
rs72890842	T	0.0673	Myopia-European	29808027	GCST006290
rs17400325	T	0.1383	Myopia-European	29808027	GCST006290
rs1550094	A	0.0861	Myopia-European	29808027	GCST006290
rs12998513	A	0.4513	Myopia-European	29808027	GCST006290
rs77016368	A	0.5618	Myopia-European	29808027	GCST006290
rs62236760	A	0.0473	Myopia-European	29808027	GCST006290
rs826222	T	0.0395	Myopia-European	29808027	GCST006290
rs2303635	C	0.8739	Myopia-European	29808027	GCST006290
rs7624084	T	0.0573	Myopia-European	29808027	GCST006290
rs7662551	A	0.056	Myopia-European	29808027	GCST006290
rs77676437	T	0.0616	Myopia-European	29808027	GCST006290
rs79504986	A	0.167	Myopia-European	29808027	GCST006290
rs5022942	A	0.0743	Myopia-European	29808027	GCST006290
rs10011267	T	0.0481	Myopia-European	29808027	GCST006290
rs411535	A	0.0429	Myopia-European	29808027	GCST006290
rs13217285	T	0.0647	Myopia-European	29808027	GCST006290
rs35909544	A	0.0625	Myopia-European	29808027	GCST006290

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs36042294	C	0.0603	Myopia-European	29808027	GCST006290
rs7754960	C	0.0559	Myopia-European	29808027	GCST006290
rs7744813	A	0.0928	Myopia-European	29808027	GCST006290
rs1064583	A	0.0467	Myopia-European	29808027	GCST006290
rs12193446	A	0.2514	Myopia-European	29808027	GCST006290
rs28613963	T	0.1783	Myopia-European	29808027	GCST006290
rs12667032	A	0.9345	Myopia-European	29808027	GCST006290
rs7829127	A	0.0915	Myopia-European	29808027	GCST006290
rs72621438	C	0.0851	Myopia-European	29808027	GCST006290
rs2272774	T	1.801	Myopia-European	29808027	GCST006290
rs7028032	T	0.045	Myopia-European	29808027	GCST006290
rs11145461	T	0.0668	Myopia-European	29808027	GCST006290
rs334354	A	0.0515	Myopia-European	29808027	GCST006290
rs11101263	T	0.047	Myopia-European	29808027	GCST006290
rs4141671	T	0.0515	Myopia-European	29808027	GCST006290
rs10824518	A	0.0597	Myopia-European	29808027	GCST006290
rs10887265	C	0.0568	Myopia-European	29808027	GCST006290
rs11202704	T	0.0454	Myopia-European	29808027	GCST006290
rs56299331	T	0.0608	Myopia-European	29808027	GCST006290
rs534311	T	0.0452	Myopia-European	29808027	GCST006290
rs11602008	A	0.1349	Myopia-European	29808027	GCST006290
rs2155413	A	0.0576	Myopia-European	29808027	GCST006290
rs715315	T	0.0486	Myopia-European	29808027	GCST006290
rs1790165	A	0.0504	Myopia-European	29808027	GCST006290
rs4768672	T	0.0478	Myopia-European	29808027	GCST006290
rs3138141	A	0.1201	Myopia-European	29808027	GCST006290
rs9508026	T	0.0451	Myopia-European	29808027	GCST006290
rs7333969	T	0.0481	Myopia-European	29808027	GCST006290
rs9513696	A	0.0558	Myopia-European	29808027	GCST006290
rs837335	C	0.04	Myopia-European	29808027	GCST006290
rs6602906	A	0.0843	Myopia-European	29808027	GCST006290
rs79501380	A	0.2733	Myopia-European	29808027	GCST006290
rs12883788	T	0.0491	Myopia-European	29808027	GCST006290
rs61991628	A	0.0608	Myopia-European	29808027	GCST006290
rs36024104	A	0.0761	Myopia-European	29808027	GCST006290
rs2738265	C	0.0525	Myopia-European	29808027	GCST006290
rs57391541	T	0.0809	Myopia-European	29808027	GCST006290
rs524952	A	0.099	Myopia-European	29808027	GCST006290

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs12898755	A	0.0533	Myopia-European	29808027	GCST006290
rs17648524	C	0.0986	Myopia-European	29808027	GCST006290
rs77409432	A	0.2585	Myopia-European	29808027	GCST006290
rs9972635	T	0.0842	Myopia-European	29808027	GCST006290
rs873693	A	0.0841	Myopia-European	29808027	GCST006290
rs80124906	A	0.6211	Myopia-European	29808027	GCST006290
rs4785742	T	0.0512	Myopia-European	29808027	GCST006290
rs2908972	A	0.0637	Myopia-European	29808027	GCST006290
rs62067167	T	0.0632	Myopia-European	29808027	GCST006290
rs3213636	A	0.1054	Myopia-European	29808027	GCST006290
rs714832	A	0.0402	Myopia-European	29808027	GCST006290
rs28488643	T	1.3391	Myopia-European	29808027	GCST006290
rs62075723	A	0.0568	Myopia-European	29808027	GCST006290
rs12965607	T	0.0678	Myopia-European	29808027	GCST006290
rs629631	T	0.1346	Myopia-European	29808027	GCST006290
rs3745548	T	0.4723	Myopia-European	29808027	GCST006290
rs7253703	A	0.512	Myopia-European	29808027	GCST006290
rs4805962	T	0.0884	Myopia-European	29808027	GCST006290
rs232660	A	0.147	Myopia-European	29808027	GCST006290
rs17274750	A	0.07	Myopia-European	29808027	GCST006290
rs2823141	A	0.046	Myopia-European	29808027	GCST006290
rs7275394	A	0.2892	Myopia-European	29808027	GCST006290
rs8132840	A	0.0502	Myopia-European	29808027	GCST006290
rs7286621	A	1.7808	Myopia-European	29808027	GCST006290
rs11913426	A	0.2963	Myopia-European	29808027	GCST006290
rs2500281	A	0.0579	Stroke-TransethnicMeta	29531354	GCST005843
rs55704954	T	0.087	Stroke-TransethnicMeta	29531354	GCST005843
rs593113	A	0.0467	Stroke-TransethnicMeta	29531354	GCST005843
rs6577389	A	-0.0457	Stroke-TransethnicMeta	29531354	GCST005843
rs12046010	A	-0.0524	Stroke-TransethnicMeta	29531354	GCST005843
rs3790607	A	-0.0761	Stroke-TransethnicMeta	29531354	GCST005843
rs2758603	T	0.0586	Stroke-TransethnicMeta	29531354	GCST005843
rs11240504	A	0.0515	Stroke-TransethnicMeta	29531354	GCST005843
rs4614977	C	0.0978	Stroke-TransethnicMeta	29531354	GCST005843
rs1516174	A	-0.0528	Stroke-TransethnicMeta	29531354	GCST005843
rs13035520	T	-0.1086	Stroke-TransethnicMeta	29531354	GCST005843
rs72889922	A	0.2591	Stroke-TransethnicMeta	29531354	GCST005843
rs12470845	A	-0.0477	Stroke-TransethnicMeta	29531354	GCST005843

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs11695863	T	-0.0485	Stroke-TransethnicMeta	29531354	GCST005843
rs6599168	T	0.059	Stroke-TransethnicMeta	29531354	GCST005843
rs17189291	T	0.0895	Stroke-TransethnicMeta	29531354	GCST005843
rs74878305	T	-0.2311	Stroke-TransethnicMeta	29531354	GCST005843
rs77737462	T	-0.0574	Stroke-TransethnicMeta	29531354	GCST005843
rs2723334	T	0.0925	Stroke-TransethnicMeta	29531354	GCST005843
rs28502066	A	-0.0576	Stroke-TransethnicMeta	29531354	GCST005843
rs6050	T	-0.0591	Stroke-TransethnicMeta	29531354	GCST005843
rs116326120	A	0.2143	Stroke-TransethnicMeta	29531354	GCST005843
rs7710854	A	-0.0677	Stroke-TransethnicMeta	29531354	GCST005843
rs11957829	A	0.0691	Stroke-TransethnicMeta	29531354	GCST005843
rs17115954	A	-0.0461	Stroke-TransethnicMeta	29531354	GCST005843
rs202808	C	0.0676	Stroke-TransethnicMeta	29531354	GCST005843
rs7750826	A	-0.0663	Stroke-TransethnicMeta	29531354	GCST005843
rs77190919	A	0.084	Stroke-TransethnicMeta	29531354	GCST005843
rs9504375	A	-0.1151	Stroke-TransethnicMeta	29531354	GCST005843
rs958762	C	0.047	Stroke-TransethnicMeta	29531354	GCST005843
rs7744623	T	0.0617	Stroke-TransethnicMeta	29531354	GCST005843
rs2223588	T	0.0475	Stroke-TransethnicMeta	29531354	GCST005843
rs3176336	A	0.0445	Stroke-TransethnicMeta	29531354	GCST005843
rs2894439	A	-0.218	Stroke-TransethnicMeta	29531354	GCST005843
rs6908662	T	0.0441	Stroke-TransethnicMeta	29531354	GCST005843
rs17705303	A	-0.0755	Stroke-TransethnicMeta	29531354	GCST005843
rs10248020	T	-0.0423	Stroke-TransethnicMeta	29531354	GCST005843
rs2158496	T	0.0747	Stroke-TransethnicMeta	29531354	GCST005843
rs2526619	A	-0.0815	Stroke-TransethnicMeta	29531354	GCST005843
rs2107595	A	0.0882	Stroke-TransethnicMeta	29531354	GCST005843
rs1029510	A	0.0452	Stroke-TransethnicMeta	29531354	GCST005843
rs2390117	A	0.1094	Stroke-TransethnicMeta	29531354	GCST005843
rs57957164	A	0.0548	Stroke-TransethnicMeta	29531354	GCST005843
rs35197511	T	-0.1168	Stroke-TransethnicMeta	29531354	GCST005843
rs42377	A	-0.0526	Stroke-TransethnicMeta	29531354	GCST005843
rs11971846	A	-0.0741	Stroke-TransethnicMeta	29531354	GCST005843
rs12675410	T	-0.0482	Stroke-TransethnicMeta	29531354	GCST005843
rs12681792	A	0.0512	Stroke-TransethnicMeta	29531354	GCST005843
rs1471859	A	-0.0471	Stroke-TransethnicMeta	29531354	GCST005843
rs2931351	T	-0.0408	Stroke-TransethnicMeta	29531354	GCST005843
rs7822041	A	-0.0418	Stroke-TransethnicMeta	29531354	GCST005843

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs2383205	A	-0.0415	Stroke-TransethnicMeta	29531354	GCST005843
rs1333048	A	-0.0539	Stroke-TransethnicMeta	29531354	GCST005843
rs10812752	A	0.0426	Stroke-TransethnicMeta	29531354	GCST005843
rs657659	A	-0.0436	Stroke-TransethnicMeta	29531354	GCST005843
rs28560429	A	0.1006	Stroke-TransethnicMeta	29531354	GCST005843
rs7077598	A	0.1202	Stroke-TransethnicMeta	29531354	GCST005843
rs56941894	T	0.1123	Stroke-TransethnicMeta	29531354	GCST005843
rs7394038	T	0.0426	Stroke-TransethnicMeta	29531354	GCST005843
rs6584579	A	0.0448	Stroke-TransethnicMeta	29531354	GCST005843
rs61007562	T	0.0907	Stroke-TransethnicMeta	29531354	GCST005843
rs77544954	A	0.0816	Stroke-TransethnicMeta	29531354	GCST005843
rs55815645	A	-0.101	Stroke-TransethnicMeta	29531354	GCST005843
rs12577661	T	-0.0478	Stroke-TransethnicMeta	29531354	GCST005843
rs55693083	A	-0.0518	Stroke-TransethnicMeta	29531354	GCST005843
rs117876107	A	0.1408	Stroke-TransethnicMeta	29531354	GCST005843
rs11603150	T	0.0452	Stroke-TransethnicMeta	29531354	GCST005843
rs9651613	A	0.0533	Stroke-TransethnicMeta	29531354	GCST005843
rs72930274	T	0.0649	Stroke-TransethnicMeta	29531354	GCST005843
rs76349035	A	-0.1235	Stroke-TransethnicMeta	29531354	GCST005843
rs470928	A	-0.0836	Stroke-TransethnicMeta	29531354	GCST005843
rs36053597	T	0.0609	Stroke-TransethnicMeta	29531354	GCST005843
rs77248876	A	0.1179	Stroke-TransethnicMeta	29531354	GCST005843
rs11064171	T	-0.1337	Stroke-TransethnicMeta	29531354	GCST005843
rs1981440	T	0.048	Stroke-TransethnicMeta	29531354	GCST005843
rs58989256	A	-0.0493	Stroke-TransethnicMeta	29531354	GCST005843
rs73338169	C	0.1215	Stroke-TransethnicMeta	29531354	GCST005843
rs10777230	A	0.0452	Stroke-TransethnicMeta	29531354	GCST005843
rs10774623	A	0.061	Stroke-TransethnicMeta	29531354	GCST005843
rs7137828	T	-0.0731	Stroke-TransethnicMeta	29531354	GCST005843
rs12427276	A	-0.0625	Stroke-TransethnicMeta	29531354	GCST005843
rs10744777	T	0.0641	Stroke-TransethnicMeta	29531354	GCST005843
rs3752631	A	0.0587	Stroke-TransethnicMeta	29531354	GCST005843
rs11066283	A	-0.0774	Stroke-TransethnicMeta	29531354	GCST005843
rs11066322	A	-0.0526	Stroke-TransethnicMeta	29531354	GCST005843
rs79280766	A	0.1518	Stroke-TransethnicMeta	29531354	GCST005843
rs10459401	T	-0.0884	Stroke-TransethnicMeta	29531354	GCST005843
rs4942561	T	0.0655	Stroke-TransethnicMeta	29531354	GCST005843
rs4942570	A	-0.0624	Stroke-TransethnicMeta	29531354	GCST005843

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs7333406	T	0.0545	Stroke-TransethnicMeta	29531354	GCST005843
rs7318338	A	0.0554	Stroke-TransethnicMeta	29531354	GCST005843
rs9564881	A	-0.0658	Stroke-TransethnicMeta	29531354	GCST005843
rs148225043	C	0.2739	Stroke-TransethnicMeta	29531354	GCST005843
rs7176568	T	0.0479	Stroke-TransethnicMeta	29531354	GCST005843
rs6496123	A	-0.0476	Stroke-TransethnicMeta	29531354	GCST005843
rs7200604	T	0.0514	Stroke-TransethnicMeta	29531354	GCST005843
rs747762	T	-0.0739	Stroke-TransethnicMeta	29531354	GCST005843
rs78884723	T	-0.1288	Stroke-TransethnicMeta	29531354	GCST005843
rs62089968	A	0.12	Stroke-TransethnicMeta	29531354	GCST005843
rs117953218	T	-0.0605	Stroke-TransethnicMeta	29531354	GCST005843
rs917016	T	-0.0516	Stroke-TransethnicMeta	29531354	GCST005843
rs878484	T	-0.048	Stroke-TransethnicMeta	29531354	GCST005843
rs72903534	A	0.1326	Stroke-TransethnicMeta	29531354	GCST005843
rs7235691	T	0.0771	Stroke-TransethnicMeta	29531354	GCST005843
rs2043304	T	-0.0522	Stroke-TransethnicMeta	29531354	GCST005843
rs60314748	T	0.0493	Stroke-TransethnicMeta	29531354	GCST005843
rs9749384	T	-0.0433	Stroke-TransethnicMeta	29531354	GCST005843
rs10402802	T	0.1011	Stroke-TransethnicMeta	29531354	GCST005843
rs6074239	T	-0.0771	Stroke-TransethnicMeta	29531354	GCST005843
rs9653750	A	0.0611	Stroke-TransethnicMeta	29531354	GCST005843
rs8184945	A	0.0438	Stroke-TransethnicMeta	29531354	GCST005843
rs12170360	A	0.0477	Stroke-TransethnicMeta	29531354	GCST005843
rs139169	A	0.0563	Stroke-TransethnicMeta	29531354	GCST005843
rs2500281	A	0.0532	Stroke-EuropeanMeta	29531354	GCST006908
rs2065524	A	-0.0778	Stroke-EuropeanMeta	29531354	GCST006908
rs55704954	T	0.0862	Stroke-EuropeanMeta	29531354	GCST006908
rs1537407	T	-0.0554	Stroke-EuropeanMeta	29531354	GCST006908
rs228725	T	0.0365	Stroke-EuropeanMeta	29531354	GCST006908
rs11121153	A	0.0405	Stroke-EuropeanMeta	29531354	GCST006908
rs284234	C	-0.0485	Stroke-EuropeanMeta	29531354	GCST006908
rs17035646	A	0.0536	Stroke-EuropeanMeta	29531354	GCST006908
rs77793475	C	0.1579	Stroke-EuropeanMeta	29531354	GCST006908
rs761291	A	0.0364	Stroke-EuropeanMeta	29531354	GCST006908
rs12057512	A	0.051	Stroke-EuropeanMeta	29531354	GCST006908
rs75178217	A	-0.0746	Stroke-EuropeanMeta	29531354	GCST006908
rs6695915	A	0.0687	Stroke-EuropeanMeta	29531354	GCST006908
rs112476957	T	-0.061	Stroke-EuropeanMeta	29531354	GCST006908

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs10922534	T	0.0534	Stroke-EuropeanMeta	29531354	GCST006908
rs12752866	T	-0.0383	Stroke-EuropeanMeta	29531354	GCST006908
rs7513849	A	0.0382	Stroke-EuropeanMeta	29531354	GCST006908
rs6537837	T	0.0432	Stroke-EuropeanMeta	29531354	GCST006908
rs1777606	A	0.0408	Stroke-EuropeanMeta	29531354	GCST006908
rs10776752	T	0.0648	Stroke-EuropeanMeta	29531354	GCST006908
rs7553733	T	-0.0393	Stroke-EuropeanMeta	29531354	GCST006908
rs909269	T	-0.0475	Stroke-EuropeanMeta	29531354	GCST006908
rs1052053	A	0.0576	Stroke-EuropeanMeta	29531354	GCST006908
rs4950874	T	-0.0355	Stroke-EuropeanMeta	29531354	GCST006908
rs11240746	A	-0.0422	Stroke-EuropeanMeta	29531354	GCST006908
rs7532642	T	0.0397	Stroke-EuropeanMeta	29531354	GCST006908
rs17015183	A	-0.0473	Stroke-EuropeanMeta	29531354	GCST006908
rs17015250	T	-0.0348	Stroke-EuropeanMeta	29531354	GCST006908
rs12476527	T	-0.0472	Stroke-EuropeanMeta	29531354	GCST006908
rs4665894	T	0.1678	Stroke-EuropeanMeta	29531354	GCST006908
rs79835740	A	0.0851	Stroke-EuropeanMeta	29531354	GCST006908
rs9309154	T	-0.0447	Stroke-EuropeanMeta	29531354	GCST006908
rs12105026	A	-0.0433	Stroke-EuropeanMeta	29531354	GCST006908
rs11647	C	-0.0465	Stroke-EuropeanMeta	29531354	GCST006908
rs79488306	A	0.0648	Stroke-EuropeanMeta	29531354	GCST006908
rs6728833	T	0.0411	Stroke-EuropeanMeta	29531354	GCST006908
rs55863310	A	-0.0901	Stroke-EuropeanMeta	29531354	GCST006908
rs13035520	T	-0.1001	Stroke-EuropeanMeta	29531354	GCST006908
rs143229390	T	0.1077	Stroke-EuropeanMeta	29531354	GCST006908
rs72864740	A	0.0532	Stroke-EuropeanMeta	29531354	GCST006908
rs301816	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs11121240	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs67551275	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs4845604	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs115045402	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs12123821	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs61816761	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs2070901	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs2056625	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs1617333	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs1102705	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs10158467	G	1	Asthma-European-Ferreira	30929738	GCST007799

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs17668708	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs12023876	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs3856439	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs74180212	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs78545931	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs60227565	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs12470864	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs72823641	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs1861245	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs2381712	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs6755248	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs10187276	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs34290285	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs35570272	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs4491851	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs1806656	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs7625643	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs62296577	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs7626218	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs9860547	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs55661102	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs11715524	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs1684466	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs190438685	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs45613035	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs17454584	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs62322662	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs16903574	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs11742240	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs7734635	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs1837253	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs1898671	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs6594499	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs6866614	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs3749833	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs2299012	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs113010607	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs449454	G	1	Asthma-European-Ferreira	30929738	GCST007799

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs139088362	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs116189786	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs28798705	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs3116989	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs28522747	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs62408233	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs58521088	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs9372120	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs55743914	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs6927172	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs12531500	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs6954667	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs57585717	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs4722758	G	1	Asthma-European-Ferreira	30929738	GCST007799
rs2221641	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs13277355	A	1	Asthma-European-Ferreira	30929738	GCST007799
rs62557312	C	1	Asthma-European-Ferreira	30929738	GCST007799
rs7848215	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs274943	T	1	Asthma-European-Ferreira	30929738	GCST007799
rs12568266	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs14271	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs114190122	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs12143327	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs74346392	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7551641	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs12074934	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7534485	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs116493700	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs73912949	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs1435547	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs13405366	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs1877101	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs12622534	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7588010	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs4143116	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs4954192	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs72974161	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76247873	A	1	Asthma-European-Yucesoy	25918132	GCST002875

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs16841200	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7576072	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76833157	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76506302	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs61741390	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs6436285	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs13391419	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs116212486	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs10174165	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs6442708	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs79143957	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs27387	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs75515191	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs3796392	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs9871967	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs190141647	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs1479279	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs6858365	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs114326390	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7693389	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs342958	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs16995986	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs111361850	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs74964132	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs77156671	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs115683773	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs2309284	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs908084	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs28375794	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs16867528	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76684306	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs16867713	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs73132886	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs74935252	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs3852186	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7733624	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7720886	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76314368	A	1	Asthma-European-Yucesoy	25918132	GCST002875

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs77885874	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7735563	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs9367895	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs2479808	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs79014439	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs16894878	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs2646928	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs116278466	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs351328	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs61613191	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs116146467	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs17167307	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs56868568	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs73302615	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs79835348	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76917448	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs10435178	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs10237103	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs10268774	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs11971779	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs73737676	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs114252942	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs6586977	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs59294057	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs16876083	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs117628011	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs55798256	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs2514805	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7000447	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs4736655	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs117405208	G	1	Asthma-European-Yucesoy	25918132	GCST002875
rs117260909	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs60244812	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs118106262	C	1	Asthma-European-Yucesoy	25918132	GCST002875
rs76780579	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs114646018	T	1	Asthma-European-Yucesoy	25918132	GCST002875
rs7020553	A	1	Asthma-European-Yucesoy	25918132	GCST002875
rs17461620	T	1	Asthma-European-Yucesoy	25918132	GCST002875

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs10157802	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs2352521	T	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs4129267	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs1348135	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs6694672	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs1122396	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs10924970	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs4268898	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs1420101	T	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs12634582	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs9870718	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs12521260	T	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs10455025	C	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs20541	A	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs7705042	A	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs1233578	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs2855812	T	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs9272346	A	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs2325291	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs10951405	T	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs10233459	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs12543811	G	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs12542922	A	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs2073617	A	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs992969	A	1	Asthma-MultiEthnic-Demenais	29273806	GCST005212
rs7542900	C	1	T2D-African	22238593	GCST001369
rs7560163	C	1	T2D-African	22238593	GCST001369
rs2722769	C	1	T2D-African	22238593	GCST001369
rs7107217	C	1	T2D-African	22238593	GCST001369
rs2237897	C	1	T2D-African	22238593	GCST001369
rs824248	T	1	T2D-African	22238593	GCST001369
rs7903146	T	1	T2D-African	22238593	GCST001369
rs335810	A	1	T2D-African	22238593	GCST001369
rs3842770	A	1	T2D-African	22238593	GCST001369
rs2283228	A	1	T2D-African	22238593	GCST001369
rs335810	A	1	T2D-African	22238593	GCST001369
rs2283228	A	1	T2D-African	22238593	GCST001369
rs12613372	G	1	T2D-African	22238593	GCST001369

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs2244020	G	1	T2D-African	22238593	GCST001369
rs17359493	G	1	T2D-African	22238593	GCST001369
rs11043007	G	1	T2D-African	22238593	GCST001369
rs2244020	G	1	T2D-African	22238593	GCST001369
rs17359493	G	1	T2D-African	22238593	GCST001369
rs679992	T	1	T2D-African	22238593	GCST001369
rs10231619	T	1	T2D-African	22238593	GCST001369
rs7003257	T	1	T2D-African	22238593	GCST001369
rs7903146	T	1	T2D-African	22238593	GCST001369
rs343092	T	1	T2D-African	22238593	GCST001369
rs7903146	T	1	T2D-African	22238593	GCST001369
rs343092	T	1	T2D-African	22238593	GCST001369
rs231356	T	1	T2D-African	22238593	GCST001369
rs7769051	A	1	T2D-African	22238593	GCST001369
rs6930576	A	1	T2D-African	22238593	GCST001369
rs773506	G	1	T2D-African	22238593	GCST001369
rs2358944	G	1	T2D-African	22238593	GCST001369
rs2106294	T	1	T2D-African	22238593	GCST001369
rs10461617	A	1	T2D-EuropeanAsian	23209189	GCST001759
rs11165354	A	1	T2D-EuropeanAsian	23209189	GCST001759
rs8050136	A	1	T2D-EuropeanAsian	23209189	GCST001759
rs6723108	T	1	T2D-EuropeanAsian	23209189	GCST001759
rs7903146	T	1	T2D-EuropeanAsian	22101970	GCST001326
rs2383208	T	1	T2D-EuropeanAsian	23209189	GCST001759
rs6134031	T	1	T2D-EuropeanAsian	26584805	GCST004782
rs17106184	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs243021	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs7612463	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4689388	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs702634	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4457053	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs329122	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs2796441	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs11787792	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs7957197	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs1359790	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs2028299	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs8042680	A	1	T2D-EuropeanAsian	28869590	GCST004894

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs8050136	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4430796	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs12970134	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs11671664	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4812829	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs17106184	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs243021	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4689388	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs702634	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4457053	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs329122	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs4607517	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs2796441	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs7957197	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs10507349	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs576674	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs1359790	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs7985179	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs8042680	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs9940149	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs8050136	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs12970134	A	1	T2D-EuropeanAsian	28869590	GCST004894
rs340874	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs2867125	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs780094	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs7578597	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs3923113	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs2943640	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs1470579	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs6808574	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs7754840	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs864745	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs12681990	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs516946	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs896854	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs10811661	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs2421016	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs1552224	C	1	T2D-EuropeanAsian	28869590	GCST004894

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs1531343	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs4275659	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs11873305	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs12454712	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs340874	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs780094	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs7578597	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs11123406	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs3923113	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs7607980	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs2943640	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs1470579	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs7754840	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs864745	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs516946	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs944801	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs10811661	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs1552224	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs1531343	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs4275659	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs2925979	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs11873305	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs1470579	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs10811661	C	1	T2D-EuropeanAsian	28869590	GCST004894
rs2296172	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs1801282	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs11708067	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs6815464	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs2706785	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs459193	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs9505118	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs3130501	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs4897182	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7041847	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs12571751	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs231362	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs10830963	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs11063069	G	1	T2D-EuropeanAsian	28869590	GCST004894

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs10507349	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs576674	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7172432	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7178572	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs9940149	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7202877	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs3786897	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs4420638	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7674212	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs2296172	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7578326	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs1801282	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs11708067	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs2706785	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs459193	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs9505118	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs3130501	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs4897182	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs12571751	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs231362	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs163184	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs10830963	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs11063069	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7178572	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7202877	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs4420638	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs7178572	G	1	T2D-EuropeanAsian	28869590	GCST004894
rs11123406	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs6813195	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2050188	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs1535500	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs622217	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs17168486	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2191349	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs9648716	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs13266634	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs13292136	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs11257655	T	1	T2D-EuropeanAsian	28869590	GCST004894

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs1111875	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs7903146	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs7111341	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2237892	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs5215	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs10842994	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs7961581	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs7985179	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2925979	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs3794991	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs10923931	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs4607103	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs6813195	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2050188	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs17168486	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2191349	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs9648716	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs13266634	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs13292136	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs11257655	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs1111875	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs7903146	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2421016	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2237892	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs5215	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs2129869	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs10842994	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs3794991	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs13266634	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs7903146	T	1	T2D-EuropeanAsian	28869590	GCST004894
rs849134	A	1	T2D-Multiethnic	27189021	GCST003619
rs7766070	A	1	T2D-Multiethnic	27189021	GCST003619
rs9687833	A	1	T2D-Multiethnic	27189021	GCST003619
rs11927381	C	1	T2D-Multiethnic	27189021	GCST003619
rs13266634	C	1	T2D-Multiethnic	27189021	GCST003619
rs6857	C	1	T2D-Multiethnic	27189021	GCST003619
rs9273401	G	1	T2D-Multiethnic	27189021	GCST003619
rs10811661	T	1	T2D-Multiethnic	27189021	GCST003619

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs3768321	T	1	T2D-Multiethnic	27189021	GCST003619
rs34872471	C	1	T2D-Multiethnic	27189021	GCST003619
rs12453394	A	1	T2D-Multiethnic	29358691	GCST005898
rs7589501	A	1	T2D-Multiethnic	29358691	GCST005898
rs1002061	A	1	T2D-Multiethnic	29358691	GCST005898
rs11708067	A	1	T2D-Multiethnic	29358691	GCST005898
rs4234733	A	1	T2D-Multiethnic	29358691	GCST005898
rs703983	A	1	T2D-Multiethnic	29358691	GCST005898
rs10404333	A	1	T2D-Multiethnic	29358691	GCST005898
rs505922	C	1	T2D-Multiethnic	29358691	GCST005898
rs66502159	C	1	T2D-Multiethnic	29358691	GCST005898
rs340874	C	1	T2D-Multiethnic	29358691	GCST005898
rs757110	C	1	T2D-Multiethnic	29358691	GCST005898
rs3794205	G	1	T2D-Multiethnic	29358691	GCST005898
rs11616380	G	1	T2D-Multiethnic	29358691	GCST005898
rs2396083	G	1	T2D-Multiethnic	29358691	GCST005898
rs12912009	G	1	T2D-Multiethnic	29358691	GCST005898
rs2358954	G	1	T2D-Multiethnic	29358691	GCST005898
rs10011174	G	1	T2D-Multiethnic	29358691	GCST005898
rs2820443	T	1	T2D-Multiethnic	29358691	GCST005898
rs10771367	T	1	T2D-Multiethnic	29358691	GCST005898
rs7113297	T	1	T2D-Multiethnic	29358691	GCST005898
rs146662075	T	1	T2D-Multiethnic	29358691	GCST005898
rs115884658	A	1	T2D-Multiethnic	29358691	GCST005898
rs6986080	G	1	T2D-Multiethnic	29358691	GCST005898
rs9502570	A	1	T2D-Multiethnic	24509480	GCST002352
rs702634	A	1	T2D-Multiethnic	24509480	GCST002352
rs6937795	A	1	T2D-Multiethnic	24509480	GCST002352
rs10788575	A	1	T2D-Multiethnic	24509480	GCST002352
rs3923113	A	1	T2D-Multiethnic	24509480	GCST002352
rs17791513	A	1	T2D-Multiethnic	24509480	GCST002352
rs1552224	A	1	T2D-Multiethnic	24509480	GCST002352
rs4812829	A	1	T2D-Multiethnic	24509480	GCST002352
rs7041847	A	1	T2D-Multiethnic	24509480	GCST002352
rs12571751	A	1	T2D-Multiethnic	24509480	GCST002352
rs12970134	A	1	T2D-Multiethnic	24509480	GCST002352
rs6813195	C	1	T2D-Multiethnic	24509480	GCST002352
rs6808574	C	1	T2D-Multiethnic	24509480	GCST002352

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs1727313	C	1	T2D-Multiethnic	24509480	GCST002352
rs10510110	C	1	T2D-Multiethnic	24509480	GCST002352
rs1561927	C	1	T2D-Multiethnic	24509480	GCST002352
rs10190052	C	1	T2D-Multiethnic	24509480	GCST002352
rs319598	C	1	T2D-Multiethnic	24509480	GCST002352
rs2820446	C	1	T2D-Multiethnic	24509480	GCST002352
rs2812533	C	1	T2D-Multiethnic	24509480	GCST002352
rs1111875	C	1	T2D-Multiethnic	24509480	GCST002352
rs5215	C	1	T2D-Multiethnic	24509480	GCST002352
rs16861329	C	1	T2D-Multiethnic	24509480	GCST002352
rs2028299	C	1	T2D-Multiethnic	24509480	GCST002352
rs9936385	C	1	T2D-Multiethnic	24509480	GCST002352
rs1801282	C	1	T2D-Multiethnic	24509480	GCST002352
rs516946	C	1	T2D-Multiethnic	24509480	GCST002352
rs10842994	C	1	T2D-Multiethnic	24509480	GCST002352
rs7163757	C	1	T2D-Multiethnic	24509480	GCST002352
rs2943640	C	1	T2D-Multiethnic	24509480	GCST002352
rs7612463	C	1	T2D-Multiethnic	24509480	GCST002352
rs17106184	G	1	T2D-Multiethnic	24509480	GCST002352
rs3132524	G	1	T2D-Multiethnic	24509480	GCST002352
rs10507349	G	1	T2D-Multiethnic	24509480	GCST002352
rs7795991	G	1	T2D-Multiethnic	24509480	GCST002352
rs4273712	G	1	T2D-Multiethnic	24509480	GCST002352
rs7756992	G	1	T2D-Multiethnic	24509480	GCST002352
rs163184	G	1	T2D-Multiethnic	24509480	GCST002352
rs849135	G	1	T2D-Multiethnic	24509480	GCST002352
rs10830963	G	1	T2D-Multiethnic	24509480	GCST002352
rs2075423	G	1	T2D-Multiethnic	24509480	GCST002352
rs8108269	G	1	T2D-Multiethnic	24509480	GCST002352
rs1359790	G	1	T2D-Multiethnic	24509480	GCST002352
rs4430796	G	1	T2D-Multiethnic	24509480	GCST002352
rs12899811	G	1	T2D-Multiethnic	24509480	GCST002352
rs12427353	G	1	T2D-Multiethnic	24509480	GCST002352
rs3802177	G	1	T2D-Multiethnic	24509480	GCST002352
rs4458523	G	1	T2D-Multiethnic	24509480	GCST002352
rs2796441	G	1	T2D-Multiethnic	24509480	GCST002352
rs7178572	G	1	T2D-Multiethnic	24509480	GCST002352
rs9472138	T	1	T2D-Multiethnic	24509480	GCST002352

Table 8 continued

rsid	EA	effect_size	Trait	PMID	Accession
rs2284219	T	1	T2D-Multiethnic	24509480	GCST002352
rs7903146	T	1	T2D-Multiethnic	24509480	GCST002352
rs11257655	T	1	T2D-Multiethnic	24509480	GCST002352
rs17168486	T	1	T2D-Multiethnic	24509480	GCST002352
rs1535500	T	1	T2D-Multiethnic	24509480	GCST002352
rs2261181	T	1	T2D-Multiethnic	24509480	GCST002352
rs4402960	T	1	T2D-Multiethnic	24509480	GCST002352
rs10811661	T	1	T2D-Multiethnic	24509480	GCST002352
rs7845219	T	1	T2D-Multiethnic	24509480	GCST002352
rs243088	T	1	T2D-Multiethnic	24509480	GCST002352
rs11717195	T	1	T2D-Multiethnic	24509480	GCST002352

Table 9. R^2 values for traits with robust ancestry correlations.

Trait	EUR_r2	AFR_r2	NAT_r2
Black vs. blond hair color	-0.537	0.547	-0.229
Black vs. red hair color	-0.509	0.508	-0.197
BMI smoking interaction	-0.493	0.521	-0.244
Breast cancer	0.435	-0.464	0.224
Cleft palate	-0.356	0.415	-0.254
Common carotid intima-media thickness in HIV infection	0.340	-0.420	0.296
Cutaneous squamous cell carcinoma	0.470	-0.414	0.098
Diisocyanate-induced asthma	-0.614	0.630	-0.268
Fear of minor pain	-0.634	0.722	-0.417
Gestational age at birth in premature rupture of membrane-initiated deliveries child effect	-0.389	0.415	-0.200
Glycemic traits pregnancy	0.373	-0.419	0.234
Heel bone mineral density	-0.494	0.567	-0.333
Hip circumference	0.445	-0.420	0.134
Ischemic stroke small artery occlusion	0.391	-0.407	0.182
Low tan response	-0.684	0.712	-0.318
Low white blood cell count	-0.739	0.822	-0.445
Metabolite levels HVA-5-HIAA Factor score	-0.389	0.437	-0.245
Methotrexate phramacokinetics acute lymphoblastic leukemia	0.457	-0.478	0.218
Migraine with aura	-0.378	0.404	-0.196
Neutrophil count in HIV-infection	-0.739	0.822	-0.445
Neutrophil level response to clozapine in treatment-resistant schizophrenia	-0.739	0.822	-0.445
Obesity-related traits	-0.612	0.749	-0.516

Table 9 continued

Trait	EUR_r2	AFR_r2	NAT_r2
Optic disc area	-0.548	0.614	-0.339
Plateletcrit	0.343	-0.428	0.310
Resistin levels	-0.394	0.448	-0.258
Response to methylphenidate treatment in attention-deficit-over-hyperactivity disorder blood pressure	-0.430	0.486	-0.275
Schizophrenia	-0.369	0.431	-0.266
Self-reported math ability	-0.653	0.720	-0.380
Skin colour saturation	-0.619	0.589	-0.193
Skin pigmentation	-0.526	0.541	-0.233
Skin sensitivity to sun	-0.567	0.530	-0.163
Supraventricular ectopy	-0.497	0.543	-0.281
Systemic lupus erythematosus	-0.511	0.521	-0.219
Tumor biomarkers	-0.400	0.433	-0.219
Type 2 diabetes nephropathy	-0.455	0.499	-0.261
White blood cell count monocyte	0.392	-0.469	0.305
Wilms tumor	-0.471	0.496	-0.230

PUBLICATIONS

1. **Chande, A.T.**, Rishishwar, L., Conley, A.B., Valderrama-Aguirre, A., Medina-Rivas, M.A. and Jordan, I.K. (2020) Ancestry effects on type 2 diabetes genetic risk inference in Hispanic/Latino populations. *BMC Medical Genetics*. In press.

2. **Chande, A.T.**, Rishishwar, L., Ban, D., Nagar, S.D., Conley, A.B., Rowell, J., Valderrama-Aguirre, A., Medina-Rivas, M.A. and Jordan, I.K. (2020) The phenotypic consequences of genetic divergence between admixed Latin American populations: Antioquia and Chocó, Colombia. *Genome Biology and Evolution*. In review.

3. **Chande, A.T.**, Wang, L., Rishishwar, L., Conley, A.B., Norris, E.T., Valderrama-Aguirre, A. and Jordan, I.K. (2018) GlobAl Distribution of GEnetic Traits (GADGET) web server: polygenic trait scores worldwide. *Nucleic Acids Res*, **46**, W121-W126.

4. **Chande, A.T.,** Rowell, J., Rishishwar, L., Conley, A.B., Norris, E.T., Valderrama-Aguirre, A., Medina-Rivas, M.A. and Jordan, I.K. (2017) Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Chocó and Antioquia. *Sci Rep*, **7**, 17127.

5. Espitia-Navarro, H.F., **Chande, A.T.**, Nagar, S.D., Smith, H., Jordan, I.K., and Rishishwar, L. (2020) STing: accurate and ultrafast genomic profiling with exact sequence matches. *Nucleic Acids Res*. In press.

6. Weitz, J.S., Harris, M., **Chande, A.T.**, Gussler, J.W., Rishishwar, L., and Jordan, I.K. (2020) Online COVID-19 Dashboard Calculates How Risky Reopenings and Gatherings Can Be. *Sci Am*.

7. Norris, E.T., Rishishwar, L., **Chande, A.T.**, Conley, A.B., Ye, K., Valderrama-Aguirre, A. and Jordan, I.K. (2020) Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol*, **21**, 29.

8. Norris, E.T., Rishishwar, L., Wang, L., Conley, A.B., **Chande, A.T.**, Dabrowski, A.M., Valderrama-Aguirre, A. and Jordan, I.K. (2019) Assortative Mating on Ancestry-Variant Traits in Admixed Latin American Populations. *Front Genet*, **10**, 359.

9. Medina-Cordoba, L.K., **Chande, A.T.**, Rishishwar, L., Mayer, L.W., Valderrama-Aguirre, L.C., Valderrama-Aguirre, A., Gaby, J.C., Kostka, J.E. and Jordan, I.K. (2019) Genomic characterization and computational phenotyping of nitrogen-fixing bacteria isolated from Colombian sugarcane fields. *bioRxiv*, 780809.

10. Espitia, H., **Chande, A.T.**, Jordan, I.K., and Rishishwar, L. (2019). US Patent App. 15/726,005.

11. Crisan, C.V., **Chande, A.T.**, Williams, K., Raghuram, V., Rishishwar, L., Steinbach, G., Watve, S.S., Yunker, P., Jordan, I.K. and Hammer, B.K. (2019) Analysis of *Vibrio cholerae* genomes identifies new type VI secretion system gene clusters. *Genome Biol*, **20**, 163.

12. Bernardy, E.E., Petit, R.A., 3rd, Moller, A.G., Blumenthal, J.A., McAdam, A.J., Priebe, G.P., **Chande, A.T.**, Rishishwar, L., Jordan, I.K., Read, T.D. *et al.* (2019) Whole-Genome Sequences of *Staphylococcus aureus* Isolates from Cystic Fibrosis Lung Infections. *Microbiol Resour Announc*, **8**.

13. Medina-Cordoba, L.K., **Chande, A.T.**, Rishishwar, L., Mayer, L.W., Marino-Ramirez, L., Valderrama-Aguirre, L.C., Valderrama-Aguirre, A., Kostka, J.E. and Jordan, I.K. (2018) Genome Sequences of 15 Klebsiella sp. Isolates from Sugarcane Fields in Colombia's Cauca Valley. *Genome Announc*, **6**.

14. Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R. and Köster, J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, **15**, 475-476.

15. Cho, C., **Chande, A.T.**, Gakhar, L., Hunt, J., Ketterer, M.R., and Apicella, M.A. (2018) Characterization of a nontypeable Haemophilus influenzae thermonuclease. *PLoS One*, **13**, e0197010.

16. Topaz, N., Mojib, N., **Chande, A.T.**, Kubanek, J. and Jordan, I.K. (2017) RampDB: a web application and database for the exploration and prediction of receptor activity modifying protein interactions. *Database (Oxford)*, **2017**.

17. Post, D.M.B., Slutter, B., Schilling, B., **Chande, A.T.**, Rasmussen, J.A., Jones, B.D., D'Souza, A.K., Reinders, L.M., Harty, J.T., Gibson, B.W., *et al.* (2017) Characterization of Inner and Outer Membrane Proteins from Francisella tularensis Strains LVS and Schu S4 and Identification of Potential Subunit Vaccine Candidates. *mBio*, **8**.

18. Post, D.M.B., Schilling, B., Reinders, L.M., D'Souza, A.K., Ketterer, M.R., Kiel, S.J., **Chande, A.T**., Apicella, M.A. and Gibson, B.W. (2017) Identification and characterization of AckA-dependent protein acetylation in Neisseria gonorrhoeae. *PLoS One*, **12**, e0179621.

19. Watve, S.S., **Chande, A.T**., Rishishwar, L., Marino-Ramirez, L., Jordan, I.K., and Hammer, B.K. (2016) Whole-Genome Sequences of 26 Vibrio cholerae Isolates. *Genome Announc*, **4**.

20. Im, S.B., Espitia, H.F., **Chande, A.T.**, Carleton, H.A., Jordan, I.K., and Rishishwar, L. Alignment-free virulence profiling of Shiga toxin-producing *Escherichia coli* for foodborne public health surveillance. In preparation.

21. Im, S.B., Gupta, S., Jain, M., **Chande, A.T.**, Carleton, H.A., Jordan, I.K., and Rishishwar, L. Predicting likelihood of PulseNet PFGE Patterns from Whole Genome Sequence. In preparation.

22. Haydek, J.P., Rishishwar, L., **Chande, A.T.**, Jordan, I.K, de Man, T.J.B., Tauxe, W.M.m Neish, E., Ward, A., Sitchenko, K.L., Woodworth, M.H., Neish, A.S., Dhere, T., and Kraft, C.S. (2020) Dissimilarities Between Luminal and Mucosal Microbiota among Fecal Microbiota Transplant Recipients. In review.

23. Wolff, B.J., Waller, J.L., Benitez, A.J., Gaines, A., Conley, A.B., Rishishwar, L., **Chande, A.T.**, Morrison, S.S., Jordan, I.K., Diaz, M.H., and Winchell, J.M. (2020) Genomic analysis of *Chlamydia psittaci* from a major zoonotic outbreak in two chicken processing plants. In preparation.

24. Diaz, M.H., Benitez, A.J., Wolff, B.J., Morrison, S.S., Fink, T., Liu, G., Gallagher, G.R., Hall, J., Gaines, A., **Chande, A.T.**, Rishishwar, L., Smole, S., Boxrud, D., and Winchell, J.M. (2020) Development and evaluation of a targeted resequencing panel for improved detection and characterization of respiratory pathogens during unexplained respiratory disease outbreaks. In preparation.

25. Wozniak, J.E, **Chande, A.T.**, Burd, E., Band, V., Satola, S., Farley, M., Jacon, J., King. I.K., Weiss, D.S. (2020). Absence of *mgrB* Mediates Fitness Cost-Free Colistin Resistance in Carbapenem-Resistant *Enterobacter cloacae*. In preparation.

REFERENCES

1. CDC, *Community health and program services (CHAPS): health disparities among racial/ethnic populations.* Atlanta, GA: US Department of Health and Human Services, 2008.

2. Leigh, J.A., M. Alvarez, and C.J. Rodriguez, *Ethnic Minorities and Coronary Heart Disease: an Update and Future Directions.* Curr Atheroscler Rep, 2016. **18**(2): p. 9.

3. Choudhry, S., et al., *Dissecting complex diseases in complex populations: asthma in latino americans.* Proc Am Thorac Soc, 2007. **4**(3): p. 226-33.

4. Polderman, T.J., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nat Genet, 2015. **47**(7): p. 702-9.

5. Meyer, D., et al., *A genomic perspective on HLA evolution.* Immunogenetics, 2018. **70**(1): p. 5-27.

6. Norris, E.T., et al., *Admixture-enabled selection for rapid adaptive evolution in the Americas.* 2019: p. 783845.

7. Ingram, C.J., et al., *Lactose digestion and the evolutionary genetics of lactase persistence.* Hum Genet, 2009. **124**(6): p. 579-91.

8. Simonson, T.S., et al., *Genetic evidence for high-altitude adaptation in Tibet.* Science, 2010. **329**(5987): p. 72-5.

9. Ilardo, M.A., et al., *Physiological and Genetic Adaptations to Diving in Sea Nomads.* Cell, 2018. **173**(3): p. 569-580 e15.

10. Albin, R.L., *Antagonistic pleiotropy, mutation accumulation, and human genetic disease.* Genetica, 1993. **91**(1-3): p. 279-86.

11. Rodriguez, J.A., et al., *Antagonistic pleiotropy and mutation accumulation influence human senescence and disease.* Nat Ecol Evol, 2017. **1**(3): p. 55.

12. Carter, A.J. and A.Q. Nguyen, *Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles.* BMC Med Genet, 2011. **12**: p. 160.

13. Plomin, R., C.M. Haworth, and O.S. Davis, *Common disorders are quantitative traits.* Nat Rev Genet, 2009. **10**(12): p. 872-8.

14. Zhu, X., R.S. Cooper, and R.C. Elston, *Linkage analysis of a complex disease through use of admixed populations.* Am J Hum Genet, 2004. **74**(6): p. 1136-53.

15. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nat Rev Genet, 2005. **6**(2): p. 95-108.

16. Popejoy, A.B. and S.M. Fullerton, *Genomics is failing on diversity.* Nature, 2016. **538**(7624): p. 161-164.

17. Zhong, Y., M.A. Perera, and E.R. Gamazon, *On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations.* Am J Hum Genet, 2019. **104**(6): p. 1097-1115.

18. Mogil, L.S., et al., *Genetic architecture of gene expression traits across diverse populations.* PLoS Genet, 2018. **14**(8): p. e1007586.

19. Shu, L., et al., *Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States.* PLoS Genet, 2017. **13**(9): p. e1007040.

20. Ellinghaus, D., et al., *Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci.* Nat Genet, 2016. **48**(5): p. 510-8.

21. Torkamani, A., N.E. Wineinger, and E.J. Topol, *The personal and clinical utility of polygenic risk scores.* Nat Rev Genet, 2018. **19**(9): p. 581-590.

22. Marigorta, U.M. and A. Navarro, *High trans-ethnic replicability of GWAS results implies common causal variants.* PLoS Genet, 2013. **9**(6): p. e1003566.

23. Kim, M.S., et al., *Genetic disease risks can be misestimated across global populations.* Genome Biol, 2018. **19**(1): p. 179.

24. Martin, A.R., et al., *Clinical use of current polygenic risk scores may exacerbate health disparities.* Nat Genet, 2019. **51**(4): p. 584-591.

25. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.

26. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.

27. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genet Epidemiol, 2008. **32**(3): p. 227-34.

28. Khoury, M.J., M.F. Iademarco, and W.T. Riley, *Precision Public Health for the Era of Precision Medicine.* American Journal of Preventive Medicine, 2016. **50**(3): p. 398-401.

29. Duncan, L., et al., *Analysis of Polygenic Score Usage and Performance in Diverse Human Populations.* bioRxiv, 2018: p. 398396.

30. Novembre, J. and N.H. Barton, *Tread Lightly Interpreting Polygenic Tests of Selection.* Genetics, 2018. **208**(4): p. 1351-1355.

31. Martin, A.R., et al., *Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.* Am J Hum Genet, 2017. **100**(4): p. 635-649.

32. Bustamante, C.D., E.G. Burchard, and F.M. De la Vega, *Genomics for the world.* Nature, 2011. **475**(7355): p. 163-5.

33. Need, A.C. and D.B. Goldstein, *Next generation disparities in human genomics: concerns and remedies.* Trends Genet, 2009. **25**(11): p. 489-94.

34. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.

35. Chatterjee, N., J. Shi, and M. Garcia-Closas, *Developing and evaluating polygenic risk prediction models for stratified disease prevention.* Nat Rev Genet, 2016. **17**(7): p. 392-406.

36. Chatterjee, N., et al., *Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies.* Nat Genet, 2013. **45**(4): p. 400-5, 405e1-3.

37. International Schizophrenia, C., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.

38. Malone, J., et al., *Modeling sample variables with an Experimental Factor Ontology.* Bioinformatics, 2010. **26**(8): p. 1112-1118.

39. Auton, A., et al., *Global distribution of genomic diversity underscores rich complex history of continental human populations.* Genome Res, 2009. **19**(5): p. 795-803.

40. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3.* Fly, 2012. **6**(2): p. 80-92.

41. Paila, U., et al., *GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations.* PLOS Computational Biology, 2013. **9**(7): p. e1003153.

42. *From the Centers for Disease Control and Prevention. Availability of Work-Related Lung Disease Surveillance Report, 1999.* JAMA, 2000. **283**(15): p. 1955.

43. McDonald, J.C., et al., *Incidence by occupation and industry of acute work related respiratory diseases in the UK, 1992-2001.* Occup Environ Med, 2005. **62**(12): p. 836-42.

44. Yucesoy, B., et al., *Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma.* Toxicol Sci, 2015. **146**(1): p. 192-201.

45. Flores, C., et al., *African ancestry is associated with asthma risk in African Americans.* PLoS One, 2012. **7**(1): p. e26807.

46. Bryant-Stephens, T., *Asthma disparities in urban environments.* J Allergy Clin Immunol, 2009. **123**(6): p. 1199-206; quiz 1207-8.

47. Iyengar, S.K., et al., *Genome-Wide Association and Trans-ethnic Meta-Analysis for Advanced Diabetic Kidney Disease: Family Investigation of Nephropathy and Diabetes (FIND).* PLoS Genet, 2015. **11**(8): p. e1005352.

48. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

49. Carlson, C.S., et al., *Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study.* PLoS Biol, 2013. **11**(9): p. e1001661.

50. Li, Y.R. and B.J. Keating, *Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations.* Genome Medicine, 2014. **6**(10): p. 91.

51. Qi, Q., et al., *Genetics of Type 2 Diabetes in U.S. Hispanic/Latino Individuals: Results From the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).* Diabetes, 2017. **66**(5): p. 1419-1425.

52. Waters, K.M., et al., *Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups.* PLoS Genet, 2010. **6**(8).

53. Chen, R., et al., *Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases.* PLoS Genet, 2012. **8**(4): p. e1002621.

54. Chande, A.T., et al., *Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Choco and Antioquia.* Sci Rep, 2017. **7**(1): p. 17127.

55. Banda, Y., et al., *Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort.* Genetics, 2015. **200**(4): p. 1285-95.

56. Burchard, E.G., et al., *The importance of race and ethnic background in biomedical research and clinical practice.* N Engl J Med, 2003. **348**(12): p. 1170-5.

57. Yudell, M., et al., *SCIENCE AND SOCIETY. Taking race out of human genetics.* Science, 2016. **351**(6273): p. 564-5.

58. Amberger, J.S., et al., *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders.* Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98.

59. Chande, A.T., et al., *GlobAl Distribution of GEnetic Traits (GADGET) web server: polygenic trait scores worldwide.* Nucleic Acids Res, 2018. **46**(W1): p. W121-W126.

60. Corona, E., et al., *Analysis of the genetic basis of disease in the context of worldwide human relationships and migration.* PLoS Genet, 2013. **9**(5): p. e1003447.

61. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.

62. Lambert, S.A., G. Abraham, and M. Inouye, *Towards clinical utility of polygenic risk scores.* Hum Mol Genet, 2019.

63. Turchin, M.C., et al., *Evidence of widespread selection on standing variation in Europe at height-associated SNPs.* Nat Genet, 2012. **44**(9): p. 1015-9.

64. Racimo, F., J.J. Berg, and J.K. Pickrell, *Detecting Polygenic Adaptation in Admixture Graphs.* Genetics, 2018. **208**(4): p. 1565-1584.

65. Berg, J.J., X. Zhang, and G. Coop, *Polygenic Adaptation has Impacted Multiple Anthropometric Traits.* bioRxiv, 2019: p. 167551.

66. Beiter, E.R., et al., *Polygenic selection underlies evolution of human brain structure and behavioral traits.* bioRxiv, 2017: p. 164707.

67. Berg, J.J. and G. Coop, *A population genetic signal of polygenic adaptation.* PLoS Genet, 2014. **10**(8): p. e1004412.

68. Winkler, C.A., G.W. Nelson, and M.W. Smith, *Admixture mapping comes of age.* Annu Rev Genomics Hum Genet, 2010. **11**: p. 65-89.

69. Jordan, I.K., L. Rishishwar, and A.B. Conley, *Native American admixture recapitulates population-specific migration and settlement of the continental United States.* PLoS Genet, 2019. **15**(9): p. e1008225.

70. Norris, E.T., et al., *Assortative Mating on Ancestry-Variant Traits in Admixed Latin American Populations.* Front Genet, 2019. **10**: p. 359.

71. Nagar, S.D., et al., *Population Pharmacogenomics for Precision Public Health in Colombia.* Front Genet, 2019. **10**: p. 241.

72. Norris, E.T., et al., *Genetic ancestry, admixture and health determinants in Latin America.* BMC Genomics, 2018. **19**(Suppl 8): p. 861.

73. Rishishwar, L., et al., *A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians.* Gene, 2015. **574**(2): p. 345-51.

74. Rishishwar, L., et al., *Ancestry, admixture and fitness in Colombian genomes.* Sci Rep, 2015. **5**: p. 12376.

75. Homburger, J.R., et al., *Genomic Insights into the Ancestry and Demographic History of South America.* PLoS Genet, 2015. **11**(12): p. e1005602.

76. Ruiz-Linares, A., et al., *Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals.* PLoS Genet, 2014. **10**(9): p. e1004572.

77. Moreno-Estrada, A., et al., *Reconstructing the population genetic history of the Caribbean.* PLoS Genet, 2013. **9**(11): p. e1003925.

78. Bryc, K., et al., *Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations.* Proc Natl Acad Sci U S A, 2010. **107 Suppl 2**: p. 8954-61.

79. Conley, A.B., et al., *A Comparative Analysis of Genetic Ancestry and Admixture in the Colombian Populations of Choco and Medellin.* G3 (Bethesda), 2017. **7**(10): p. 3435-3447.

80. Medina-Rivas, M.A., et al., *Choco, Colombia: a hotspot of human biodiversity.* Rev Biodivers Neotrop, 2016. **6**(1): p. 45-54.

81. Chande, A.T., et al., *Ancestry effects on type 2 diabetes genetic risk inference in Hispanic/Latino populations.* . BMC Genomics, 2020. **In press**.

82. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

83. Reich, D., et al., *Reconstructing Native American population history.* Nature, 2012. **488**(7411): p. 370-4.

84. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.

85. Delaneau, O., et al., *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.* Nat Commun, 2014. **5**: p. 3934.

86. Delaneau, O., et al., *Haplotype estimation using sequencing reads.* Am J Hum Genet, 2013. **93**(4): p. 687-96.

87. Alexander, D.H., J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res, 2009. **19**(9): p. 1655-64.

88. Weir, B.S. and C.C. Cockerham, *Estimating F-Statistics for the Analysis of Population Structure.* Evolution, 1984. **38**(6): p. 1358-1370.

89. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-D1012.

90. Malone, J., et al., *Modeling sample variables with an Experimental Factor Ontology.* Bioinformatics, 2010. **26**(8): p. 1112-8.

91. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-9.

92. Howie, B., J. Marchini, and M. Stephens, *Genotype imputation with thousands of genomes.* G3 (Bethesda), 2011. **1**(6): p. 457-70.

93. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.

94. Uribe Vélez, A., et al., *Censo General 2005* 2006, Bogotá: Departamento Administrativo Nacional de Estadística (DANE).

95. Alvarez, M.C., *Encuesta nacional de la situación nutricional en Colombia*. 2006, Bogotá: Instituto Colombiano de Bienestar Familiar.

96. Edge, M.D. and N.A. Rosenberg, *A General Model of the Relationship between the Apportionment of Human Genetic Diversity and the Apportionment of Human Phenotypic Diversity.* Hum Biol, 2015. **87**(4): p. 313-337.

97. Edge, M.D. and N.A. Rosenberg, *Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity.* Stud Hist Philos Biol Biomed Sci, 2015. **52**: p. 32-45.

98. Nyenhuis, S.M., et al., *Race is associated with differences in airway inflammation in patients with asthma.* J Allergy Clin Immunol, 2017. **140**(1): p. 257-265 e11.

99. Moorman, J.E., et al., *National surveillance for asthma--United States, 1980-2004.* MMWR Surveill Summ, 2007. **56**(8): p. 1-54.

100. Bibbins-Domingo, K., et al., *Racial differences in incident heart failure among young adults.* N Engl J Med, 2009. **360**(12): p. 1179-90.

101. Bahrami, H., et al., *Differences in the incidence of congestive heart failure by ethnicity: the multi-ethnic study of atherosclerosis.* Arch Intern Med, 2008. **168**(19): p. 2138-45.

102. Park, S.C. and Y.T. Jeen, *Genetic Studies of Inflammatory Bowel Disease-Focusing on Asian Patients.* Cells, 2019. **8**(5).

103. Nguyen, G.C., C.A. Chong, and R.Y. Chong, *National estimates of the burden of inflammatory bowel disease among racial and ethnic groups in the United States.* J Crohns Colitis, 2014. **8**(4): p. 288-95.

104. Tishkoff, S.A., et al., *Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance.* Science, 2001. **293**(5529): p. 455-62.

105. Shriner, D. and C.N. Rotimi, *Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase.* Am J Hum Genet, 2018. **102**(4): p. 547-556.

106. Yao, S., et al., *Genetic ancestry and population differences in levels of inflammatory cytokines in women: Role for evolutionary selection and environmental factors.* PLoS Genet, 2018. **14**(6): p. e1007368.

107. Zweifler, R.M., et al., *Impact of race and ethnicity on ischemic stroke. The University of California at San Diego Stroke Data Bank.* Stroke, 1995. **26**(2): p. 245-8.

108. Mahal, B.A., et al., *Prostate Cancer-Specific Mortality Across Gleason Scores in Black vs Nonblack Men.* JAMA, 2018. **320**(23): p. 2479-2481.

109. Toles, C.A., *Black men are dying from prostate cancer.* ABNF J, 2008. **19**(3): p. 92-5.

110. Kaze, A.D., et al., *Burden of chronic kidney disease on the African continent: a systematic review and meta-analysis.* BMC Nephrol, 2018. **19**(1): p. 125.

111. Crews, D.C., et al., *Poverty, race, and CKD in a racially and socioeconomically diverse urban population.* Am J Kidney Dis, 2010. **55**(6): p. 992-1000.

112. Weiss, D.J., et al., *Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000-17: a spatial and temporal modelling study.* Lancet, 2019. **394**(10195): p. 322-331.

113. Battle, K.E., et al., *Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000-17: a spatial and temporal modelling study.* Lancet, 2019. **394**(10195): p. 332-343.

114. Nosten, F.H. and A.P. Phyo, *New malaria maps.* Lancet, 2019. **394**(10195): p. 278-279.

115. Lewontin, R.C., *The apportionment of human diversity*, in *Evolutionary Biology*, T.H. Dobzhansky, M.K. Hecht, and W.C. Steere, Editors. 1972, Springer: New York, NY. p. 381-398.

116. Li, J.Z., et al., *Worldwide human relationships inferred from genome-wide patterns of variation.* Science, 2008. **319**(5866): p. 1100-4.

117. Duncan, L., et al., *Analysis of polygenic risk score usage and performance in diverse human populations.* Nat Commun, 2019. **10**(1): p. 3328.

118. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nat Genet, 2018. **50**(9): p. 1219-1224.

119. Elliott, J., et al., *Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease.* JAMA, 2020. **323**(7): p. 636-645.

120. De La Vega, F.M. and C.D. Bustamante, *Polygenic risk scores: a biased prediction?* Genome Med, 2018. **10**(1): p. 100.

121. Anauati, M.V., S. Galiani, and F. Weinschelbaum, *The rise of noncommunicable diseases in Latin America and the Caribbean: challenges for public health policies.* Lat Am Econ Rev, 2015. **24**(1): p. 11.

122. Anderson, G.F., et al., *Non-communicable chronic diseases in Latin America and the Caribbean*. 2009, Baltimore: Johns Hopkins University.

123. Casas, J.A., J.N. Dachs, and A. Bambas, *Health disparities in Latin America and the Caribbean: the role of social and economic determinants.* Equity and Health, 2001. **8**: p. 22-49.

124. Almeida-Filho, N., et al., *Research on health inequalities in Latin America and the Caribbean: bibliometric analysis (1971–2000) and descriptive content analysis (1971–1995).* American Journal of Public Health, 2003. **93**(12): p. 2037-2043.

125. Knowler, W.C., et al., *Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota.* Am J Epidemiol, 1978. **108**(6): p. 497-505.

126. Burrows, N.R., et al., *Prevalence of diabetes among Native Americans and Alaska Natives, 1990-1997: an increasing burden.* Diabetes Care, 2000. **23**(12): p. 1786-90.

127. Brancati, F.L., et al., *Incident type 2 diabetes mellitus in African American and white adults: the Atherosclerosis Risk in Communities Study.* JAMA, 2000. **283**(17): p. 2253-9.

128. Cowie, C.C., et al., *Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population: National Health And Nutrition Examination Survey 1999-2002.* Diabetes Care, 2006. **29**(6): p. 1263-8.

129. Maskarinec, G., et al., *Diabetes prevalence and body mass index differ by ethnicity: the Multiethnic Cohort.* Ethn Dis, 2009. **19**(1): p. 49-55.

130. Cowie, C.C., et al., *Full accounting of diabetes and pre-diabetes in the U.S. population in 1988-1994 and 2005-2006.* Diabetes Care, 2009. **32**(2): p. 287-94.

131. Chakraborty, R., et al., *Relationship of prevalence of non-insulin-dependent diabetes mellitus to Amerindian admixture in the Mexican Americans of San Antonio, Texas.* Genet Epidemiol, 1986. **3**(6): p. 435-54.

132. Cheng, C.Y., et al., *African ancestry and its correlation to type 2 diabetes in African Americans: a genetic admixture analysis in three U.S. population cohorts.* PLoS One, 2012. **7**(3): p. e32840.

133. Campbell, D.D., et al., *Amerind ancestry, socioeconomic status and the genetics of type 2 diabetes in a Colombian population.* PLoS One, 2012. **7**(4): p. e33570.

134. Gardner, L.I., Jr., et al., *Prevalence of diabetes in Mexican Americans. Relationship to percent of gene pool derived from native American sources.* Diabetes, 1984. **33**(1): p. 86-92.

135. Signorello, L.B., et al., *Comparing diabetes prevalence between African Americans and Whites of similar socioeconomic status.* Am J Public Health, 2007. **97**(12): p. 2260-7.

136. Robbins, J.M., et al., *Excess type 2 diabetes in African-American women and men aged 40-74 and socioeconomic status: evidence from the Third National Health and Nutrition Examination Survey.* J Epidemiol Community Health, 2000. **54**(11): p. 839-45.

137. Link, C.L. and J.B. McKinlay, *Disparities in the prevalence of diabetes: is it race/ethnicity or socioeconomic status? Results from the Boston Area Community Health (BACH) survey.* Ethn Dis, 2009. **19**(3): p. 288-92.

138. Jordan, I.K., *The Columbian Exchange as a source of adaptive introgression in human populations.* Biol Direct, 2016. **11**(1): p. 17.

139. Conley, A.B., et al., *A Comparative Analysis of Genetic Ancestry and Admixture in the Colombian Populations of Choco and Medellin.* G3 (Bethesda), 2017.

140. Hernández Romero, A., *La visibilización estadística de los grupos étnicos colombianos*. 2005, Bogotá: Departamento Administrativo Nacional de Estadística (DANE).

141. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

142. Team, R.D.C., *R: A language and environment for statistical computing.* 2008, Vienna: R Foundation for Statistical Computing.

143. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

144. Delaneau, O., et al., *Haplotype Estimation Using Sequencing Reads.* The American Journal of Human Genetics. **93**(4): p. 687-696.

145. Delaneau, O. and J. Marchini, *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.* Nature Communications, 2014. **5**: p. 3934.

146. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-959.

147. Howie, B., J. Marchini, and M. Stephens, *Genotype Imputation with Thousands of Genomes.* G3: Genes|Genomes|Genetics, 2011. **1**(6): p. 457.

148. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.

149. Viechtbauer, W., *Conducting meta-analyses in R with the metafor package.* J Stat Softw, 2010. **36**(3): p. 1-48.

150. Wray, N.R., M.E. Goddard, and P.M. Visscher, *Prediction of individual genetic risk to disease from genome-wide association studies.* Genome Res, 2007. **17**(10): p. 1520-8.

151. Dudbridge, F., *Power and predictive accuracy of polygenic risk scores.* PLoS Genet, 2013. **9**(3): p. e1003348.

152. Krapohl, E., et al., *Phenome-wide analysis of genome-wide polygenic scores.* Mol Psychiatry, 2016. **21**(9): p. 1188-93.

153. Wang, X., et al., *Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction.* J Diabetes, 2016. **8**(1): p. 24-35.

154. Vassy, J.L., et al., *Polygenic type 2 diabetes prediction at the limit of common variant detection.* Diabetes, 2014. **63**(6): p. 2172-82.

155. Talmud, P.J., et al., *Sixty-five common genetic variants and prediction of type 2 diabetes.* Diabetes, 2015. **64**(5): p. 1830-40.

156. Agardh, E., et al., *Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis.* Int J Epidemiol, 2011. **40**(3): p. 804-18.

157. Robbins, J.M., et al., *Socioeconomic status and type 2 diabetes in African American and non-Hispanic white women and men: evidence from the Third National Health and Nutrition Examination Survey.* Am J Public Health, 2001. **91**(1): p. 76-83.

158. Thompson, F.E., et al., *Interrelationships of added sugars intake, socioeconomic status, and race/ethnicity in adults in the United States: National Health Interview Survey, 2005.* J Am Diet Assoc, 2009. **109**(8): p. 1376-83.

159. Jimeno, M., M.L. Sotomayor, and L.M. Valderrama, *Chocó: diversidad cultural y medio ambiente*. 1995, Bogotá: Fondo FEN Colombia.

160. Wade, P., *Blackness and race mixture: the dynamics of racial identity in Colombia*. 1995, Balitmore: JHU Press.

161. Fagua-Duarte, J.C., et al., *Estudio nacional de consumo de sustancias psicoactivas en Colombia 2013*. 2014, Bogotá: Observatorio de Drogas de Colombia.

162. Rasouli, B., et al., *Alcohol consumption is associated with reduced risk of Type 2 diabetes and autoimmune diabetes in adults: results from the Nord-Trondelag health study.* Diabet Med, 2013. **30**(1): p. 56-64.

163. Koppes, L.L., et al., *Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies.* Diabetes Care, 2005. **28**(3): p. 719-25.

164. Joosten, M.M., et al., *Changes in alcohol consumption and subsequent risk of type 2 diabetes in men.* Diabetes, 2011. **60**(1): p. 74-9.

165. Jaddoe, V.W., et al., *Fetal exposure to parental smoking and the risk of type 2 diabetes in adult women.* Diabetes Care, 2014. **37**(11): p. 2966-73.

166. Piatti, P., et al., *Smoking is associated with impaired glucose regulation and a decrease in insulin sensitivity and the disposition index in first-degree relatives of type 2 diabetes subjects independently of the presence of metabolic syndrome.* Acta Diabetol, 2014. **51**(5): p. 793-9.

167. Yeh, H.C., et al., *Smoking, smoking cessation, and risk for type 2 diabetes mellitus: a cohort study.* Ann Intern Med, 2010. **152**(1): p. 10-7.

168. Zimmet, P., K.G. Alberti, and J. Shaw, *Global and societal implications of the diabetes epidemic.* Nature, 2001. **414**(6865): p. 782-7.

169. Zimmet, P., *Globalization, coca-colonization and the chronic disease epidemic: can the Doomsday scenario be averted?* J Intern Med, 2000. **247**(3): p. 301-10.

170. Anjana, R.M., et al., *Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR-INDIAB population-based cross-sectional study.* Lancet Diabetes Endocrinol, 2017. **5**(8): p. 585-596.

171. Zimmet, P.Z., *Diabetes and its drivers: the largest epidemic in human history?* Clin Diabetes Endocrinol, 2017. **3**: p. 1.

172. van Dieren, S., et al., *The global burden of diabetes and its complications: an emerging pandemic.* Eur J Cardiovasc Prev Rehabil, 2010. **17 Suppl 1**: p. S3-8.

173. Herman, W.H. and P. Zimmet, *Type 2 diabetes: an epidemic requiring global attention and urgent action.* Diabetes Care, 2012. **35**(5): p. 943-4.

174. *IDF Diabetes Atlas, 8th Edition*. 2019 03/06/2019]; Available from: http://www.diabetesatlas.org/.

175. Spanakis, E.K. and S.H. Golden, *Race/ethnic difference in diabetes and diabetic complications.* Curr Diab Rep, 2013. **13**(6): p. 814-23.

176. Cusi, K. and G.L. Ocampo, *Unmet needs in Hispanic/Latino patients with type 2 diabetes mellitus.* Am J Med, 2011. **124**(10 Suppl): p. S2-9.

177. Meigs, J.B., L.A. Cupples, and P.W. Wilson, *Parental transmission of type 2 diabetes: the Framingham Offspring Study.* Diabetes, 2000. **49**(12): p. 2201-2207.

178. Poulsen, P., et al., *Heritability of Type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study.* Diabetologia, 1999. **42**: p. 139-145.

179. Willemsen, G., et al., *The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium.* Twin Res Hum Genet, 2015. **18**(6): p. 762-71.

180. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nature Genetics, 2018. **50**(9): p. 1219-1224.

181. McMahon, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Research, 2018. **47**(D1): p. D1005-D1012.

182. Chande, A.T., et al., *GlobAl Distribution of GEnetic Traits (GADGET) web server: polygenic trait scores worldwide.* Nucleic Acids Research, 2018. **46**(W1): p. W121-W126.

183. Rosenberg, N.A., et al., *Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences.* Evol Med Public Health, 2019. **2019**(1): p. 26-34.

184. Marigorta, U.M., et al., *Replicability and Prediction: Lessons and Challenges from GWAS.* Trends in Genetics, 2018. **34**(7): p. 504-517.

185. Martin, A.R., et al., *Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.* The American Journal of Human Genetics, 2017. **100**(4): p. 635-649.

186. Mora, G.C., *Making Hispanics: How activists, bureaucrats, and media constructed a new American*. 2014: University of Chicago Press.

187. Wang, S., et al., *Geographic patterns of genome admixture in Latin American Mestizos.* PLoS Genet, 2008. **4**(3): p. e1000037.

188. Conley, A.B., et al., *A Comparative Analysis of Genetic Ancestry and Admixture in the Colombian Populations of Chocó and Medellín.* G3 (Bethesda, Md.), 2017. **7**(10): p. 3435-3447.

189. Bank, T.W. *The World Bank Diabetes Prevalence*. [cited 2018 12]; Available from: https://data.worldbank.org/indicator/SH.STA.DIAB.ZS.

190. Association, A.D. *Statistics About Diabetes*. [cited 2018 12/17]; Available from: http://www.diabetes.org/diabetes-basics/statistics/.

191. Health, U.D.o. *Complete Health Indicator Report of Diabetes Prevalence*. Available from: https://ibis.health.utah.gov/indicator/complete_profile/DiabPrev.html

192. Health, C.o.L.A.P., *Trends in Diabetes: Time for Action*. 2012.

193. Morales, J., et al., *A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog.* Genome biology, 2018. **19**(1): p. 21-21.

194. Morris, A.P., et al., *Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes.* Nature genetics, 2012. **44**(9): p. 981-990.

195. Cho, Y.S., et al., *Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians.* Nature Genetics, 2011. **44**: p. 67.

196. The Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**: p. 68.

197. Medina-Rivas, M.A., et al., *Chocó, Colombia: a hotspot of human biodiversity.* Revista biodiversidad neotropical, 2016. **6**(1): p. 45-54.

198. Delaneau, O., et al., *Haplotype estimation using sequencing reads.* American journal of human genetics, 2013. **93**(4): p. 687-696.

199. Delaneau, O., et al., *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.* Nature communications, 2014. **5**: p. 3934-3934.

200. Chande, A.T., et al., *Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Chocó and Antioquia.* Scientific reports, 2017. **7**(1): p. 17127.

201. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* GigaScience, 2015. **4**: p. 7-7.

202. Vilhjálmsson, Bjarni J., et al., *Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.* The American Journal of Human Genetics, 2015. **97**(4): p. 576-592.

203. Alexander, D.H., J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals.* Genome research, 2009. **19**(9): p. 1655-1664.

204. Brancati, F.L., et al., *Incident Type 2 Diabetes Mellitus in African American and White AdultsThe Atherosclerosis Risk in Communities Study.* JAMA, 2000. **283**(17): p. 2253-2259.

205. Burrows, N.R., et al., *Prevalence of diabetes among Native Americans and Alaska Natives, 1990-1997: an increasing burden.* Diabetes Care, 2000. **23**(12): p. 1786-1790.

206. Maskarinec, G., et al., *Diabetes prevalence and body mass index differ by ethnicity: the Multiethnic Cohort.* Ethnicity & disease, 2009. **19**(1): p. 49-55.

207. Chacon-Duque, J.C., et al., *Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance.* Nat Commun, 2018. **9**(1): p. 5388.

208. Moreno-Estrada, A., et al., *Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits.* Science, 2014. **344**(6189): p. 1280-5.

209. Lachance, J. and S.A. Tishkoff, *SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it.* Bioessays, 2013. **35**(9): p. 780-6.

210. Grinde, K.E., et al., *Generalizing polygenic risk scores from Europeans to Hispanics/Latinos.* Genetic Epidemiology, 2019. **43**(1): p. 50-62.

211. Ho, D.S.W., et al., *Machine Learning SNP Based Prediction for Precision Medicine.* Front Genet, 2019. **10**: p. 267.

212. Márquez-Luna, C., P.-R. Loh, and A.L. Price, *Multiethnic polygenic risk scores improve risk prediction in diverse populations.* Genetic Epidemiology, 2017. **41**(8): p. 811-823.

213. Mak, T.S.H., et al., *Polygenic scores via penalized regression on summary statistics.* Genet Epidemiol, 2017. **41**(6): p. 469-480.

214. Euesden, J., C.M. Lewis, and P.F. O'Reilly, *PRSice: Polygenic Risk Score software.* Bioinformatics, 2015. **31**(9): p. 1466-8.